국립국어원 2019-01-23

발간등록번호
11-1371028-000776-01

형태 분석 말뭉치 구축

연구 책임자

김 일 환

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '형태 분석 말뭉치 구축'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2019년 7월 ~ 2019년 12월

2019년 12월 9일

연구 책임자: 김일환(성신여자대학교)

연구 기관 성신여자대학교 산학협력단
서울대학교 산학협력단
고려대학교 산학협력단
연구 책임자 김일환
공동 연구원 박진호, 송상헌, 유현조, 윤태진, 이규범, 이도길, 정성훈, 정연주, 최운호

국문 초록

형태 분석 말뭉치 구축

이 사업은 인공지능 발전을 위한 우리말 기초 자원으로 활용될 고품질의 한국어 형태 분석 말뭉치를 구축하고, 형태 분석 말뭉치 구축을 위한 표준적인 지침을 개발하는 데 주요 목적이 있다.

사업의 범위는 크게 두 부분으로 나눌 수 있다. 첫째는 형태 분석 말뭉치 구축 지침 수립으로, <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 언어 현실에 맞게 수정하고 보다 구체화하였다. 둘째는 형태 분석 말뭉치 구축으로, 형태 분석 말뭉치 구축 지침을 바탕으로 총 300만 어절 규모(문어 200만 어절, 구어 100만 어절)의 형태 분석 말뭉치를 구축하였다.

○ 형태 분석 말뭉치 구축 지침 수립

형태 분석 말뭉치 구축 지침을 수립하기 위하여 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 기반으로 삼되 문제점을 보완하였다. 이 과정에서 한국정보통신기술협회(TTA)의 '표준 형태소 태그셋(TTAK.KO-11.0010/R1)', <물결21>의 '형태 분석 지침'을 참조하였고, 형태 분석 질의응답 게시판을 운영하여 자주 질문되는 문제를 해결할 수 있는 설명과 사례를 지침에 추가하였다. 지침 보완의 방향성과 주요 보완 사항은 다음과 같다.

① 분석 지침의 구체화

- 조사 결합형을 분석하는 기준을 명시하였다.
- 접두사, 접미사를 언제 분리하는지에 대한 지침을 명시하였다.

- 언중의 직관을 고려하여 고유명사의 범위를 보다 넓히면서도, 작업의 일관성을 위하여 구체적으로 범위를 명시하고 다양한 사례를 제시하였다.

② 언어학적 엄밀성의 추구

- '있다'의 동사 용법과 형용사 용법을 구분하는 기준을 제시하였다.
- 종결어미와 연결어미를 후행하는 문장 부호에 따라 구분하기보다 기능에 따라 구분하도록 하였다.

③ 과도한 분석 지양

- 한국어에서 유의미하게 사용되는 언어 단위를 분석 대상으로 삼기 위하여 외국어 처리 지침과 기타 기호가 포함된 어절의 처리 지침을 재정비하였다.

또한 본 사업에서 마련한 지침은 문어뿐 아니라 구어 분석 시에도 적용할 수 있도록 구성되었다. 즉 문어와 구어는 원칙적으로 동일한 형태 표지 목록을 바탕으로 동일한 방식으로 분석된다. 다만 구어는 문어와 달리 다양한 준말과 형태 변이 현상을 보여 주므로 구어에서 나타나는 준말과 형태 변이 현상의 처리 방법을 따로 명시할 필요가 있다. 또한 구어 전사 시에 이용된 특별한 마크업과 표지가 있기에, 그것의 처리 방법도 별도로 제시할 필요가 있다. 이에 지침의 말미에 구어 분석 시의 유의점에 대한 지침을 붙였는데, 구어에서 나타나는 준말과 형태 변이 현상을 되도록 분석에 반영하는 방향으로 지침의 내용을 마련하였다.

○ 300만 어절 규모의 형태 분석 말뭉치 구축

이 사업에서 구축한 형태 분석 말뭉치의 총 규모는 300만 어절(문어 200만 어절, 구어 100만 어절)이다. 형태 분석 말뭉치의 기반이 된 원시 말뭉치는 2004년 이후 생산된 현대 국어 자료를 수집한 <2018년 국어 말뭉치 연구 및 구축> 사업 결과물의 일부로서, 문어 말뭉치는 전체 신문으로 구성되어 있으며 구어 말뭉치는 공적 독백, 공적 대화, 사적 대화를 포함한다. 원시 말뭉치의 각 어절을 대상으로 형태를 분리하고 형태 분류 표

지(세분류 47종)를 부착하는 작업을 하였는데, 형태 분리의 기준이 되는 단위는 기본적으로 <우리말샘>에 등재된 단어이되, 생산성이 비교적 높은 접사도 분리하는 것을 원칙으로 삼았다.

형태 분석 말뭉치 구축은 형태 분석 지침 수립 → 분석 도구(워크벤치) 구현 → 작업자 교육 → 자동 형태소 분석 → 분석 오류 수정 → 최종 결과물 산출의 순으로 이루어졌다.

이 중 분석 오류 수정은 3단계로 이루어졌다. 1단계는 작업자 2인이 동일한 원시 말뭉치에 대한 자동 형태 분석 결과를 각자 수정하는 단계이다. 2단계는 작업자 2인의 오류 수정 결과를 비교하며 검수자 1인이 형태 분석 결과를 검수하는 단계이다. 3단계는 전체 작업 결과물에 대해 상위 작업자 그룹이 형태 결합 오류 목록, 어절 분석 중의성 목록 등을 검토하며 오류를 수정하는 단계이다.

본 사업에서는 말뭉치 구축의 편의를 도모하고 정확성을 높이기 위하여 높은 분석 정확률을 갖춘 형태소 분석기(서울대 형태소 분석기)를 사용하였다. 서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다.

한편으로 형태 분석 말뭉치 구축에 최적화된 워크벤치를 개발하였다. 워크벤치에서는 서울대 형태소 분석기의 어절 분석 결과를 보여주되 그것을 손쉽게 수정할 수 있게 하였고, 드롭다운 선택 방식 및 오류 검사를 통해 입력 오류를 원천적으로 차단하였다. 또한 동일 어절에 대한 작업자 2인의 분석 결과를 비교하고 분석 결과가 일치하지 않는 경우 그 중 옳은 분석을 선택할 수 있도록 하여 분석 결과 검수의 효율을 높였다.

차례

제1장 서론

1. 사업의 목적

○ 형태 분석 말뭉치 구축 지침 수립
○ 형태 분석 말뭉치(300만 어절) 구축

이 사업은 인공지능 발전을 위한 우리말 기초 자원으로 활용될 고품질의 한국어 형태 분석 말뭉치를 구축하고, 형태 분석 말뭉치 구축을 위한 표준적인 지침을 개발하는 데 주요 목적이 있다.

한국어 처리를 전제로 하는 인공지능 기술의 발전을 위해서는 높은 정확성을 갖춘 대규모의 형태 분석 말뭉치가 요구된다. 형태 분석은 어휘의미 분석, 구문 분석, 의미역 분석 등 상위 수준의 언어 분석에 기반이 되므로 언어처리 인공지능 기술의 발전을 위해서는 가장 기초 기술인 형태 분석기의 성능 향상이 무엇보다 중요하며, 이를 위해서는 높은 정확성과 일관성을 갖춘 대규모 학습 자원이 필요한 것이다.

그러나 2007년 배포된 <21세기 세종계획> 말뭉치 이후 공공 자원으로서의 말뭉치가 구축되지 못해 학계, 산업계 등에서 활용할 수 있는 형태 분석 말뭉치의 규모가 심각하게 부족한 상황에 처하게 되었다. 또한 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침에는 기술의 구체성이 낮은 부분이 포함되어 있어 다수의 작업자들이 지침을 기반으로 일관된 작업을 하기에는 부족한 점이 있었고, 이에 따라 구축된 말뭉치에도 분석의 일관성이 결여된 부분이 포함되어 있었다.

이에 본 사업에서는 <21세기 세종계획>의 말뭉치 구축 지침을 수정·보완한 표준적인 지침을 마련하고 그것을 바탕으로 분석의 일관성을 높인 300만 어절 규모의 형태 분

석 말뭉치를 구축하였으며, 이를 통해 공공 자원으로서의 대규모 형태 분석 말뭉치 구축의 기반을 다시금 마련하고자 하였다.

2. 사업의 범위

사업의 범위는 크게 두 부분으로 나눌 수 있다. 첫째는 형태 분석 말뭉치 구축 지침 수립으로, <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 언어 현실에 맞게 수정하고 보다 구체화하였다. 둘째는 형태 분석 말뭉치 구축으로, 형태 분석 말뭉치 구축 지침을 바탕으로 총 300만 어절 규모(문어 200만 어절, 구어 100만 어절)의 형태 분석 말뭉치를 구축하였다.

2.1. 형태 분석 말뭉치 구축 지침 수립

형태 분석 말뭉치 구축 지침을 수립하기 위하여, <21세기 세종계획> 형태 분석 말뭉치의 분석 지침을 기반으로 삼되 다른 지침과 비교・검토하며 문제점을 보완하였다.

<21세기 세종계획>의 형태 분석 말뭉치 구축 지침은 국내의 다른 형태 분석 말뭉치 구축 과정에서 두루 활용되고 있을 만큼 표준적인 지위를 확보하고 있다. 그러나 이를 바탕으로 실제 구축된 형태 분석 말뭉치에는 여러 오류가 포함되어 있으며, 이는 지침의 기술에 세밀하지 못한 부분이 있었음을 보여 준다.

이에 본 사업에서는 분석 지침의 항목별 설명을 좀 더 상세히 제시하고 다양한 사례를 포함함으로써 실제 구축 과정에서 도출되는 오류와 비일관성을 줄이고자 하였다. 이 과정에서 <물결21>의 '형태 분석 지침', 한국정보통신기술협회(TTA)의 '표준 형태소 태그셋(TTAK.KO-11.0010/R1)'을 참조하였고, 형태 분석 질의응답 게시판을 운영하여 자주 질문되는 문제를 해결할 수 있는 설명과 사례를 지침에 추가하였다.

2.2. 형태 분석 말뭉치 구축

수립된 형태 분석 말뭉치 구축 지침을 바탕으로 300만 어절 규모의 형태 분석 말뭉치를 구축하였다. 말뭉치는 문어 200만 어절, 구어 100만 어절로 구성되어 있으며 하나의 형태 분석 말뭉치 구축 지침으로 문어와 구어를 아울러 분석할 수 있음을 확인하였다.

말뭉치 구축의 편의를 도모하고 정확성을 높이기 위하여 높은 분석 정확률을 갖춘 형태소 분석기(서울대 형태소 분석기)를 사용하였다. 서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다. 한편으로 형태 분석 말뭉치 구축에 최적화된 워크벤치를 개발하였다. 워크벤치에서는 서울대 형태소 분석기의 어절 분석 결과를 보여 주되 그것을 손쉽게 수정할 수 있게 하였고, 오류 검사를 통해 입력 오류를 원천적으로 차단하였다. 또한 동일 어절에 대한 작업자 2인의 분석 결과를 비교하고 분석 결과가 일치하지 않는 경우 그중 옳은 분석을 선택할 수 있도록 하여 분석 결과 검수의 효율을 높였다.

이를 바탕으로 원시 말뭉치의 각 어절을 대상으로 형태를 분리하고 형태 분류 표지(세분류 47종)를 부착하는 작업을 하였다. 형태 분리의 기준이 되는 단위는 기본적으로 <우리말샘>에 등재된 단어이되, 생산성이 비교적 높은 접사도 분리하는 것을 원칙으로 삼았다.

제2장 형태 분석 말뭉치의 구성 및 구축 절차

1. 형태 분석 말뭉치의 구성

이 사업에서 구축한 형태 분석 말뭉치의 총 규모는 300만 어절(문어 200만 어절, 구어 100만 어절)이다. 형태 분석 말뭉치의 기반이 된 원시 말뭉치는 2004년 이후 생산된 현대 국어 자료를 수집한 <2018년 국어 말뭉치 연구 및 구축> 사업 결과물의 일부로서, 문어 말뭉치는 전체 신문으로 구성되어 있으며 구어 말뭉치는 공적 독백, 공적 대화, 사적 대화를 포함한다.

문어 말뭉치의 신문사별, 기사 주제별 분포는 다음과 같다. 2009년~2017년에 작성된 동아일보, 조선일보, 한겨레의 기사를 두루 포함하고 있다.

	2009년	2010년	2011년	2012년	2013년	2014년	2015년	2016년	2017년	합계
동아일보	63,255	65,395	68,339	99,455	70,978	102,223	65,855	96,257	65,658	697,415
조선일보	94,124	98,183	94,644	3,886	107,244	4,181	98,291	392	105,934	606,879
한겨레	72,357	72,244	72,649	85,397	73,356	87,460	73,967	86,092	72,233	695,755
합계	229,736	235,822	235,632	188,738	251,578	193,864	238,113	182,741	243,825	2,000,049

<표 1> 문어 말뭉치(200만 어절)의 구성 (신문사별)

	동아일보	조선일보	한겨레	비율	합계
경제	96,237	45,603	123,893	13.3%	265,733
과학	10,406	358	15,995	1.3%	26,759
국제	45,338	56,231	82,917	9.2%	184,486
기획	57,833	3,600	0	3.1%	61,433
문화	86,670	79,006	77,602	12.2%	243,278
사람들	7,000	33,042	2,209	2.1%	42,251
사회	102,134	105,932	170,881	18.9%	378,947
스포츠	66,340	56,537	46,321	8.5%	169,198
오피니언	60,667	46,126	28,815	6.8%	135,608
정치	141,546	124,396	130,317	19.8%	396,259
지역	23,244	56,048	16,805	4.8%	96,097
합계	697,415	606,879	695,755	100%	2,000,049

<표 2> 문어 말뭉치(200만 어절)의 구성 (기사 주제별)

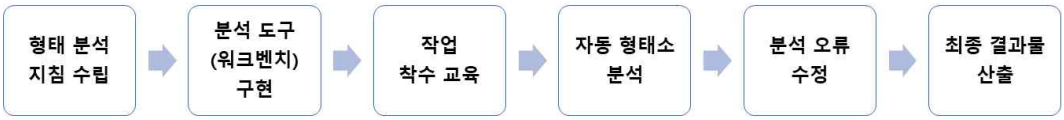
구어 말뭉치의 구성은 다음과 같다. 공적 독백 장르에 속하는 자료로는 뉴스, 강연, 학술 발표 자료가 포함되었고, 공적 대화 장르에 속하는 자료로는 방송 대화와 인터뷰 자료가 포함되었으며, 사적 대화 장르에 속하는 자료로는 2인의 일상대화 자료가 포함되었다.

분류	자료명	비고	어절 수	비율
공적 독백	EBS 정오 뉴스	2018년	24,756	2.5%
	강연	학술발표회(한글대장경의 구성과 번역의 문제) 학술특강(사회언어학 이후의 사회언어학) 학술발표회(언어 연구와 문법) 학술발표회(언어 연구와 정보, 복합지식) 학술발표회(학습용 기본 명사 연어 빈도 사전)	48,665	4.9%
	제1~5회 일송학술대회	2009년~2013년	62,153	6.2%
공적 대화	EBS 초대석		203,015	20.3%
	김어준의 뉴스공장	TBS	130,031	13%
	뜨거운 사이다	온스타일, 2017년	68,891	6.9%
	인터뷰 전사	정현종, 윤여순, 정바비 인터뷰 (2014년)	19,485	1.9%
	최고의 요리 비결	EBS	150,395	15%
	팟캐스트	프로파일러 배상훈의 CRIME (2016~2018년) 풀어 듣는 문화 이야기 (2016~2017년) 서늘한 마음썰 (2018년)	60,215	6%
사적 대화	2인 일상대화	2018년	233,854	23.3%
합계			1,001,460	100%

<표 3> 구어 말뭉치(100만 어절)의 구성

2. 형태 분석 말뭉치 구축 절차

형태 분석 말뭉치 구축 절차는 다음과 같다.



2.1. 형태 분석 지침 수립

형태 분석 말뭉치 구축의 첫 번째 단계는 기존 형태 분석 말뭉치 구축 지침의 미비점을 분석하고 보완 방안을 마련하는 것이다.

기존 형태 분석 말뭉치 구축 지침 중 가장 대표적인 것은 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침이다. 이 지침은 국내의 형태 분석 말뭉치 구축 과정에 두루 활용되고 있을 만큼 표준적인 지위를 확보하고 있으며, 특히 분석 표지(tag set)는 일부 표지를 제외하고는 여러 형태 분석 말뭉치에 공통적으로 적용되고 있어 말뭉치의 활용을 용이하게 하였다.

그럼에도 이 지침에는 보완이 필요한 부분이 포함되어 있는데, 주요 보완 사항은 다음과 같다.

○ 분석 지침의 구체성이 부족하고 충분한 예시가 제시되지 않은 경우가 있다.
○ 개별 언어 요소에 부여하는 분석 표지가 적절하지 않은 경우가 있다.

'형태 분석 지침 수립' 단계에서는 이러한 점을 보완하여 형태 분석의 타당성과 신뢰성을 높일 수 있는 방안을 마련하고자 하였다.

형태 분석 지침 수립 단계는 형태 분석 작업을 시작하기 전에 우선적으로 수행되어야

할 단계이지만, 이를 바탕으로 분석 오류를 수정하는 단계에서 발생하는 문제들을 반영하면서 반복적으로 수행되어야 할 단계이기도 하다. 본 사업에서는 질의응답 게시판을 운영하여 형태 분석 작업 시 지침으로 해결되기 어려운 부분에 대한 질문을 상시 수합하였으며, 그러한 문제를 해결할 수 있도록 지속적으로 지침을 수정하였다.

이러한 과정을 통해 마련된 최종 지침의 구체적인 내용과 주요 보완 사항은 제3장에서 보일 것이다.

2.2. 분석 도구(워크벤치) 구현

형태 분석 말뭉치 구축의 두 번째 단계는 형태 분석 오류를 효율적으로 수정하고 작업 진도를 모니터링할 수 있는 분석 도구를 구현하는 것이다.

본 사업에서는 형태 분석 오류를 효율적으로 수정하고 작업 진도를 관리하기 위해 웹 기반의 워크벤치를 구축, 활용하였다. 웹 기반의 워크벤치는 작업자들의 동시 접속과 다중 작업 수행을 돕고 <우리말샘>과 연동하여 사전을 쉽게 참조할 수 있게 하는 등 효율적인 수정 작업을 지원하였다. <그림 1>은 작업자 화면에서 '학기제'라는 단어에 마우스 포인터를 두었을 때 단어 아래에 돋보기 모양의 아이콘이 생기는 것을 보여 주는데, 이 아이콘을 클릭하면 <그림 2>와 같이 별도 창이 열려 <우리말샘>에서 '학기제'를 검색한 결과를 보여 준다.

<그림 1> 워크벤치 작업자 화면과 <우리말샘>의 연동(1)

<그림 2> 워크벤치 작업자 화면과 <우리말샘>의 연동(2)

워크벤치의 홈 화면에는 최신 지침과 작업 시 유의 사항을 제시하여 모든 작업자가 공통된 지침을 확인할 수 있도록 하였다.

<그림 3> 워크벤치의 홈 화면

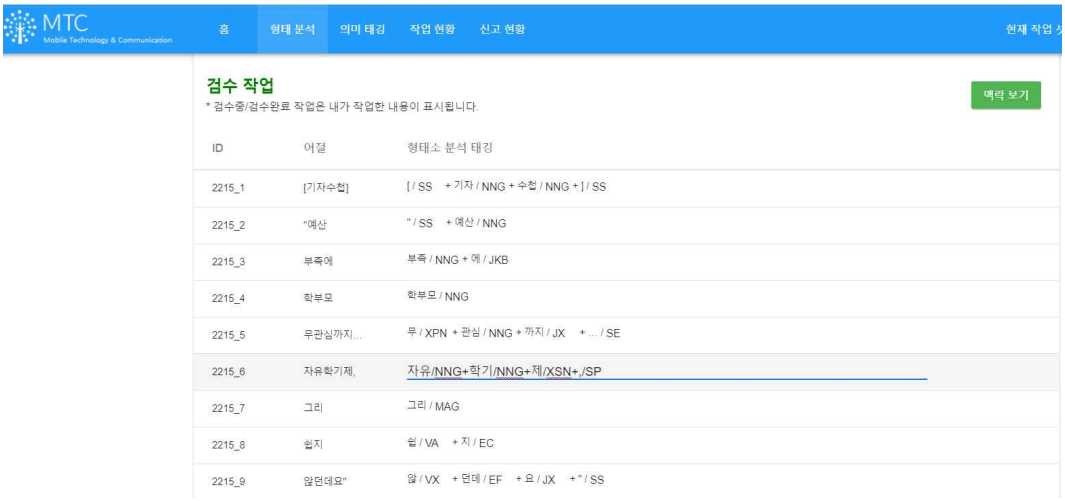
또한 워크벤치 사용자의 역할에 따라 최적화된 기능을 활용할 수 있도록 하였다. 워크벤치 사용자는 작업자, 검수자, 운영자로 나뉘며, 각 부류의 사용자는 담당하는 역할이 다른 만큼 워크벤치에서 활용할 기능도 서로 다르다.

먼저 작업자는 자신이 맡은 작업 분량에 대하여 서울대 형태소 분석기의 어절 분석 결과를 확인하고 그것에서 보이는 오류를 수정하는 역할을 한다. 이에 워크벤치는 작업자들이 맡은 분량에 대하여 서울대 형태소 분석기의 어절 분석 결과를 화면에서 실시간으로 보여주는 역할을 하였으며, 그 분석 결과를 드롭다운 선택 방식 또는 직접 수정 방식으로 수정할 수 있도록 지원하였다. 분석 결과 수정 시 의도치 않은 오류가 발생하는 것을 원천적으로 차단하기 위하여 형태 표지를 수정해야 할 경우에는 드롭다운 메뉴를 통해 수정하도록 하였고, 분석 방식을 수정해야 할 경우에는 직접 입력하여 수정할 수 있게 하되, 형식에 대한 유효성 제약을 두어 잘못된 방식으로 수정되었을 경우 오류 메시지를 산출하도록 하였다.

아래의 <그림 4>는 워크벤치의 작업자 화면에서 형태 표지 부분을 클릭하면 드롭다

운 메뉴가 나오는 것을 보여 준다. <그림 5>는 형태 표지뿐 아니라 분석 방식을 수정할 필요가 있을 때 해당 라인을 더블클릭하여 분석 방식을 직접 수정할 수 있음을 보여 준다. 이때 만약 분석 결과에 형식 오류가 포함되면 오류 메시지가 나오면서 작업 결과를 제출할 수 없도록 하였다.

<그림 4> 작업자 화면의 드롭다운 메뉴

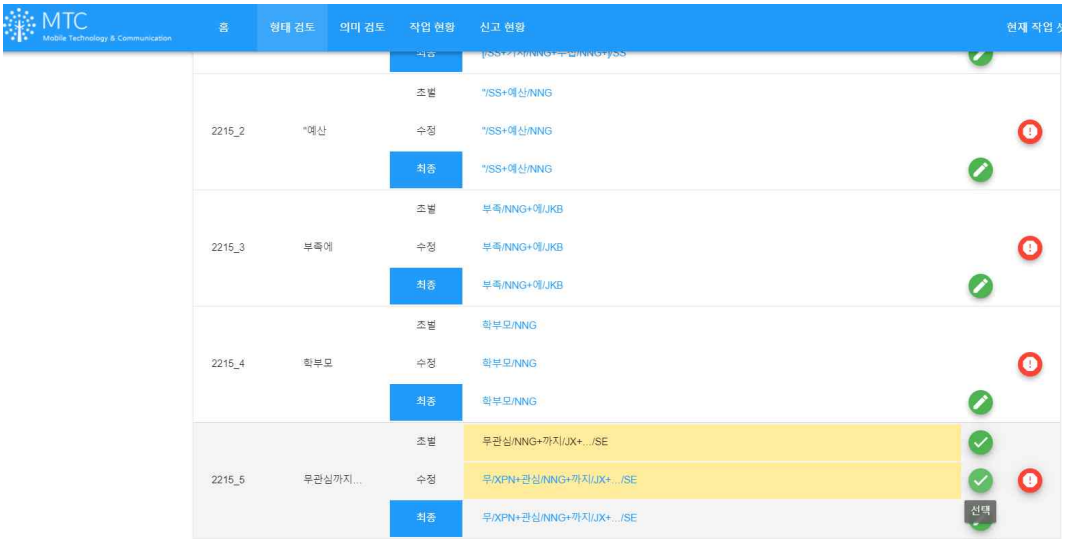


<그림 5> 작업자 화면에서의 형태 분석 결과 직접 수정

다음으로 검수자는 작업자들의 형태 분석 수정 결과를 검토하고 그것에서 보이는 오류를 수정하는 역할을 한다. 작업자 2인이 원시 말뭉치의 동일한 부분을 각자 수정하였으므로 검수자는 작업자 2인의 수정 결과를 비교하며 올바른 분석을 최종적으로 선택하게 된다. 이러한 작업을 용이하게 하기 위하여 워크벤치는 동일 어절에 대한 2인 작업자의 수정 결과를 비교하여 보여주었고, 그 중 올바른 분석을 선택하는 방식으로 손쉽게 검수할 수 있도록 하였다. 2인 작업자의 수정 결과 모두에 오류가 있을 때에는 검수자가 직접 입력하여 분석 결과를 수정할 수 있게 하되, 역시 형식에 대한 유효성 제약을 두어 잘못된 방식으로 수정되었을 경우 오류 메시지를 산출하도록 하였다.

<그림 6>은 워크벤치의 검수자 화면으로, 각 어절에 대한 작업자 2인의 형태 분석 수정 결과를 비교하여 보여주고 있다. 작업자 2인의 형태 분석 수정 결과가 일치하지 않는 경우에 대해서는 해당 부분을 노란색으로 표시하여 강조하였다. 검수자는 두 분석 결과 중 올바른 분석을 클릭하여 최종 결과로 반영할 수 있다. 만약 두 분석 결과 모두에 오류가 있다면 '최종' 라인을 더블클릭하여 직접 최종 결과를 수정할 수 있는데, 이때 만약 분석 결과에 형식 오류가 포함되면 오류 메시지가 나오면서 작업 결과를 제출할 수

없도록 하였다.

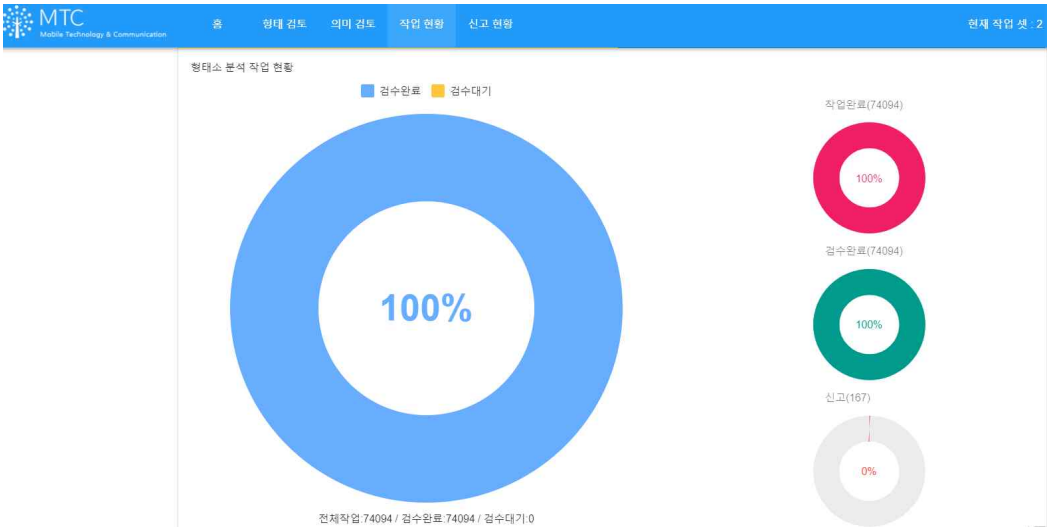


<그림 6> 워크벤치의 검수자 화면

작업자, 검수자의 작업 중 질문이나 논의가 필요한 사안이 생기는 경우에는 각 어절 옆에 있는 신고 버튼을 눌러 해당 어절에 대한 질문·논의 사안을 기록할 수 있게 하였다. <그림 6>에서 보이는 빨간색의 느낌표 아이콘이 신고 버튼에 해당한다.

마지막으로 운영자는 작업자, 검수자의 작업 진행 상황을 모니터링하고 형태 분석 시 발생하는 여러 문제 및 질문 사안을 해결하는 역할을 한다. 이에 워크벤치에 '작업 현황' 탭을 두어 전체 작업자·검수자의 작업 진척 정도 및 개별 작업자·검수자의 작업 진척 정도를 확인할 수 있게 하였다. 또한 '신고 현황' 탭을 두어 작업자, 검수자가 제기한 질문 및 논의 사안을 확인하고, 각 사안에 대한 해결 방법을 입력하여 모든 작업자, 검수자가 그 내용을 확인할 수 있게 하였다. '신고 현황' 탭은 작업자·검수자와 운영자 사이의 질의응답 게시판의 역할을 수행하였으며, 여기에서 확인된 질의응답 내용을 바탕으로 형태 분석 말뭉치 구축 지침을 지속적으로 수정하고 보완하였다. <그림 7>은 작업 현황

화면을, <그림 8>은 신고 현황 화면을 보인 것이다.



<그림 7> 작업 현황 화면

<그림 8> 신고 현황 화면

이와 같이 작업자, 검수자, 운영자 각자의 업무를 돕는 웹 기반의 워크벤치를 구현하여 작업의 효율성을 도모하였다. 워크벤치의 기능 및 화면 구성은 형태 분석 작업 진행 과정 중 작업자, 검수자, 운영자의 요청에 따라 지속적으로 수정되고 보완되었다.

2.3. 작업 착수 교육

형태 분석 말뭉치 구축의 세 번째 단계는 형태 분석 말뭉치 구축 지침과 분석 도구 사용에 대한 교육을 실시하는 것이다. 아울러 비밀 유지와 자료 보안, 문서 보안 등과 관련한 보안 교육도 필요하다. 이에 사업 전체 참여자를 대상으로 보안 교육 및 형태 분석 말뭉치 구축 지침, 워크벤치 사용에 대한 교육을 실시하였다.

특히 형태 분석 말뭉치 구축 지침에 대한 교육은 1회의 교육만으로는 불충분하며 수시로 개별 작업자의 작업 수행에 대한 피드백이 이루어져야 한다. 본 사업의 형태 분석 작업은 2인의 작업자가 수행한 형태 분석 수정 결과를 1인의 검수자가 검토하는 방식으로 이루어졌으므로, 개별 작업자가 빈번히 만들어 내는 오류를 검수자가 파악하여 지속적으로 피드백하였다. 또한 전체 분석 결과의 오류와 일관성을 검토하는 최종 검수 단계에서 작업자들이 공통적으로 만들어 내는 오류를 파악하여 수시 교육을 실시하였다.

2.4. 자동 형태소 분석

형태 분석 말뭉치 구축의 네 번째 단계는 자동 형태소 분석이다. 작업자는 자동 형태소 분석 결과를 토대로 오류를 수정하게 되며, 따라서 높은 분석 정확률을 갖춘 자동 형태 분석 도구를 사용하는 것이 작업의 효율을 높이기 위한 관건이 된다. 이에 본 사업에서는 높은 분석 정확률을 갖춘 자동 형태 분석 도구(서울대 형태소 분석기)를 사용하였다.

서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를

철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다. 딥러닝(특히 LSTM)의 sequence tagging 알고리즘을 적용하였고, 형태소 분절과 품사 태그 부착의 두 단계로 나누어 각 어절을 처리한다. 딥러닝 적용을 위해 형태소 분절 문제를 한글 음절의 유형 분류 문제로 변형하여 접근하였는데, 이를 위해 1200만 어절 형태의미분석 말뭉치 전체에 대한 검토를 통해 총망라적 목록으로서 한글 음절의 200개 유형을 정리하였다.

한편 Mecab 고유명사 사전의 30여만 개 표제어에 대해 다른 품사와 중의성을 야기할 수 있는 문제 단어들을 수작업으로 제외하여 고유명사 사전을 구축하였고, <우리말샘>의 복합명사 중 하이픈(-)으로 연결된 것과 ^으로 연결된 것을 각각 사전으로 구축하여 복합명사의 분절 여부 판단에 활용하였다. 하이픈으로 연결된 것도 중의성 야기 우려가 있는 것들은 수작업으로 검토하여 제외하였다.

이와 같은 과정을 통해 개발된 서울대 형태소 분석기는 문어 자료에 대해 98%의 F1 score를 보였다. 딥러닝 기반의 형태소 분석기이므로, 딥러닝의 훈련 자료가 추가되면 될수록 형태소 분석기의 성능은 높아진다. 이에 본 사업에서는 형태 분석 말뭉치 중간 산출물을 추가 학습 자료로 삼아 형태소 분석기를 다시금 업그레이드하였고, 이를 적용함으로써 형태 분석 작업의 효율성을 더욱 높였다.

이에 더해 본 사업의 대상이 되는 300만 어절 말뭉치에 3번 이상 등장하는 어절을 추출하여, 그 중 형태 분석의 중의성이 없는 어절을 대상으로 정확한 형태 분석 결과를 대응시키는 '기분석 어절 사전'을 구축하였다. 이것으로 서울대 형태소 분석기의 분석 결과를 보완함으로써 작업의 효율성을 더욱 높일 수 있었다. <그림 9>는 '기분석 어절 사전'에 포함된 내용의 일부를 보인 것이다.

38057	다뤄졌다.	다루/VV+어/EC+지/VX+었/EP+다/EF+./SF
38058	다뤄지지	다루/VV+어/EC+지/VX+지/EC
38059	다뤄질	다루/VV+어/EC+지/VX+ㄹ/ETM
38060	다뤘기	다루/VV+었/EP+기/ETN
38061	다뤘다.	다루/VV+었/EP+다/EF+./SF
38062	다뤘던	다루/VV+었/EP+던/ETM
38063	다뤘을	다루/VV+었/EP+을/ETM
38064	다르게	다르/VA+게/EC
38065	다르고	다르/VA+고/EC
38066	다르고,	다르/VA+고/EC+,/SP
38067	다르기	다르/VA+기/ETN
38068	다르냐고	다르/VA+냐고/EC
38069	다르니	다르/VA+니/EC
38070	다르다"고	다르/VA+다/EF+"/SS+고/JKQ
38071	다르다"며	다르/VA+다/EF+"/SS+며/EC
38072	다르다.	다르/VA+다/EF+./SF
38073	다르다”고	다르/VA+다/EF+”/SS+고/JKQ
38074	다르다”며	다르/VA+다/EF+”/SS+며/EC

<그림 9> 기분석 어절 사전

2.5. 분석 오류 수정

서울대 형태소 분석기의 어절 분석 결과 및 기분석 어절 사전의 적용 결과는 웹 기반 워크벤치의 작업자 화면에서 확인되며, 작업자는 이 분석 결과의 오류를 수정하는 방식으로 형태 분석 말뭉치 구축 작업을 진행하였다.

작업자들은 3인 1조(작업자 2인-검수자 1인)를 이루어 형태 분석의 오류를 수정하였으며, 오류 수정은 다음과 같이 3단계로 이루어졌다.

○ 1단계: 작업자 2인의 자동 형태 분석 오류 수정

각 조는 오류를 수정할 원시 말뭉치를 분배받는다. 조에 할당된 원시 말뭉치의 각 어절에 대한 자동 형태 분석 결과가 작업자 2인의 작업 화면에 제시된다. 이를 토대로 작업자 2인은 동일 원시 말뭉치에 대해 각자 오류 수정 작업을 진행한다.

○ 2단계: 작업자 2인의 오류 수정 결과에 대한 검수자 1인의 검수

검수자 1인이 작업자 2인의 오류 수정 결과를 검토하면서, 작업자 2인의 오류 수정 결과가 서로 불일치하는 경우에 더욱 주의하여 형태 분석의 오류를 수정한다.

○ 3단계: 전체 작업 결과물에 대한 최종 검수

전체 조의 형태 분석 결과 검수가 끝나면, 공동연구원을 중심으로 한 상위 그룹이 전체 형태 분석 말뭉치에 대하여 형태 결합 방식의 오류(예: 체언에 어미가 결합하는 것으로 분석된 오류)나 원어절과 분석 결과 사이의 자소 불일치 오류가 있는지 검토하여 수정한다. 또한 동일한 형식의 어절을 둘 이상의 방식으로 분석한 어절 중의성 목록을 검토하여 진정한 중의성과 분석 오류로 인한 중의성을 구별하고, 후자에 대해서는 분석 중의성이 발생하지 않도록 수정한다. 또한 형태 분석 지침의 변경 이력에 따라 변경된 지침의 적용 여부를 검토하여 수정한다.

이 중 3단계 최종 검수 과정에 대해 자세히 보이기로 한다. 전체 작업 결과물에는 지침에 대한 숙지 또는 이해의 차이, 개별 작업자의 단순 실수 등으로 인한 오류가 포함되어 있다. 우선 형태 결합 방식에 대한 메타 지식을 토대로 알고리즘을 구성하여 아래와 같은 결합 오류 유형을 추출하고 수정하였다.

2391-26431	수은주는	수은주/NNG+는/ETM
2391-56058	지을	짓/VV+을/JKO
2391-77358	만들것”	만들/VV+것/NNB+”/SS
2391-96797	깨질거라곤	깨지/VV+ㄹ/ETM+거/NNB+라고/EC+ㄴ/JX
2391-103957	적은	적/VA+은/JX
2391-203088	초라도	초/NNB+라도/EC
2391-262786	뜬(던)지고	뜬/NA+(/SS+던/NA+)/SS+지/NA+고/EC
2391-338655	명확지	명확/NNG+지/EC

<그림 10> 형태 결합 오류

위는 명사+어미, 동사+조사, 형용사+조사, 동사+명사와 같이 문법적으로 올바르지 않은 결합으로 분석된 어절이 있음을 보여 준다. 이처럼 문법적으로 불가능한 방식으로 분석된 어절을 추출하는 알고리즘을 구성하여 형태 결합 오류를 포함한 어절을 자동으로

추출하고 수정하였다. 특히 위의 예에서 '만들것', '깨질거라곤', '명확지'처럼 생략된 요소를 복원하여 분석해야 하는 어절에서 생략 요소를 복원하지 않은 오류가 발생하기 쉬운데, 형태 결합 오류 어절을 추출하는 알고리즘을 통하여 효과적으로 오류 어절을 발견할 수 있다.

아래는 동일한 형식의 어절을 둘 이상의 방식으로 분석한 어절 중의성 목록의 예를 보인 것이다. 첫 번째 열에는 원어절, 두 번째 열과 네 번째 열에는 원어절의 분석 결과, 세 번째 열과 다섯 번째 열에는 각 분석의 빈도가 제시되어 있다. 이 중에는 일반명사 또는 고유명사로 분석된 '아이티'처럼 실제로 중의성이 있는 어절도 있지만, '아이스링크'처럼 지침에 따르면 '아이스링크/NNG'로 분석되어야 할 것이 '아이스/NNG+링크/NNG'로도 분석된, 분석 오류로 인한 중의성도 있다. 이 중 분석 오류로 인한 중의성은 올바른 분석으로 수정함으로써 중의성 없이 하나의 분석 방식으로 통일되도록 하였다.

아이스링크	아이스/NNG+링크/NNG	2	아이스링크/NNG	1
아이스크림	아이스크림/NNG	19	아이스크림/NNP	1
아이엠에프	아이엠에프/NNG	1	아이엠에프/NNP	1
아이티	아이티/NNG	6	아이티/NNP	44
아이팟	아이팟/NNG	1	아이팟/NNP	2
아이패드	아이패드/NNG	3	아이패드/NNP	15
아주대병원	아주대/NNP+병원/NNG	1	아주대병원/NNP	2
아찔하	아찔/MAG+하/XSA	2	아찔하/VA	4

<그림 11> 어절 분석 중의성 검토 자료

또한 형태 표지별 목록도 추출하여 잘못된 형태 표지가 부여된 요소가 없는지를 검토하고 수정하였다. 아래는 SL(외국어) 형태 표지가 부여된 요소의 목록과 빈도를 추출한 자료의 일부인데, '㎞'처럼 SW(기타 기호) 표지가 부여되어야 할 요소가 SL로 잘못 분석되어 있음을 볼 수 있다. 이처럼 형태 표지별 목록에서 발견되는 오류도 수정하였다.

```
119 seoul               1
120 spark               1
121 stops               1
122 tax                 1
123 teplestay           1
124 the                 1
125 twitter             1
126 twitterkr           1
127 unmerciful          1
128 up                  1
129 wibro               1
130 yozm                1
131 km                  1
```

<그림 12> 형태 표지별 목록(SL)

본 사업에서는 위와 같은 자동 형태 분석과 3단계의 오류 수정 공정을 하나의 회기로 묶어 총 3회기를 진행하였다.

회기	기간	분석 말뭉치 구축량	
		회기당	누적
1회기	9월 16일~10월 15일	문어: 90만	문어: 90만
2회기	10월 7일~11월 3일	문어: 90만	문어: 180만
3회기	10월 29일~12월 6일	문어: 20만 구어: 100만	문어: 200만 구어: 100만

<표 4> 회기별 기간 및 구축량

2.6. 최종 결과물 산출

형태 분석 말뭉치 구축의 최종 단계는 오류가 수정된 형태 분석 결과물을 JSON 형식으로 변환하여 최종 결과물을 산출하는 것이다. JSON 형식의 기본 구조와 예시는 부록에서 제시하였다.

제3장 형태 분석 말뭉치 구축 지침 수립

1. 지침 수립 과정

지침 수립 과정은 <21세기 세종계획> 형태 분석 말뭉치 구축 지침의 미비점을 검토하는 단계와, 다른 지침을 참고하며 그것을 보완하는 단계로 나뉜다.

<21세기 세종계획>의 형태 분석 말뭉치 구축 지침은 국내의 다른 형태 분석 말뭉치 구축 과정에서도 활용되고 있을 만큼 표준적인 지위를 확보하고 있으며, 특히 분석 표지(tag set)는 일부 표지를 제외하고는 여러 형태 분석 말뭉치에 공통적으로 적용되고 있어 말뭉치의 활용을 용이하게 하였다.

그러나 이를 바탕으로 실제 구축된 형태 분석 말뭉치에는 여러 오류가 포함되어 있으며, 이는 기존 지침에 세밀하지 못한 부분이 있었음을 보여 준다.

이에 본 사업에서는 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 기반으로 하되, 일부 내용을 수정하고 항목별 설명을 상세화하며 다양한 사례를 포함함으로써 실제 구축 과정에서 도출되는 오류를 줄이고자 하였다.

1.1. <21세기 세종계획> 형태 분석 말뭉치 구축 지침 검토

<21세기 세종계획>이 형태 분석 말뭉치 구축 지침에는 보완이 필요한 부분이 포함되어 있는데, 주요 보완 사항은 다음과 같다.

○ 분석 지침의 구체성이 부족하고 충분한 예시가 제시되지 않은 경우가 있다.
○ 개별 언어 요소에 부여하는 분석 표지가 적절하지 않은 경우가 있다.

1.1.1. 분석 지침이 구체화되어야 하는 경우

① 고유명사(NNP)

고유명사 중 '지명' 지침을 보면, 아래와 같은 내용이 제시되어 있음을 확인할 수 있다. 그런데 지명의 종류로 제시된 것이 제한적이어서 '관동팔경, 경인공업지대, 가자지구, 경포해수욕장, 광화문광장' 등 일반적으로 지명으로 인식되는 명칭을 고유명사로 처리할 것인지 아닌지를 지침만으로는 판단하기 어렵다. '광화문광장'은 시설물이나 구조물로 볼 수도 있는데, '건축물이나 시설물 혹은 구조물의 이름' 지침도 이러한 사례를 포함하고 있지 않다. 이에 따라 작업자마다 어떤 경우에는 고유명사로 처리하고 어떤 경우에는 일반명사로 처리하는 경우가 발생하게 된다.

(2) 지명

(가) 내륙, 바다, 강, 산, 산맥, 호수, 섬, 만, 계곡, 늪, 주 등의 이름
예: 카스피해/NNP, 템즈강/NNP, 태백산맥/NNP, 미시시피호/NNP, 네바다주/NNP

(나) 도(道), 시(市), 읍(邑), 면(面), 리(里), 군(郡), 구(區), 동(洞), 골, 촌 등의 이름은 그 구역의 종류를 나타내는 말과 함께 전체가 고유명사가 된다.
예: 서울특별시/NNP, 성북구/NNP, 강진군/NNP, 인창동/NNP, 빨래골/NNP, 해방촌/NNP

(4) 건축물이나 시설물 혹은 구조물의 이름

(가) 도로, 항만, 철도, 전철, 지하철 및 그 명칭과 함께 쓰이는 부대시설은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.
예: 부산항/NNP, 대전역/NNP, 서울지하철/NNP, 테헤란로/NNP

(나) 빌딩, 박물관, 극장 등 건물명은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된

다.
예: 서울역사/NNP, 세종문화회관/NNP, 개나리유치원/NNP, 고려대학교/NNP
국립중앙박물관/NNP, 국립민속박물관/NNP, 구텐베르크박물관/NNP
신라호텔/NNP, 미도파백화점/NNP, 동궁예식장/NNP, 명보극장/NNP, 고대병원/NNP

② 동사(VV), 형용사(VA)

동사와 형용사 지침은 아래와 같이 간단하게 제시되어 있다. 하지만 한국어는 형용사가 동사와 유사한 언어로서 동사와 형용사 중 어떤 형태 표지를 부여해야 할지 결정하기 어려운 요소가 있으며, 특히 용언 '있다'는 빈도도 매우 높은 데다 동사와 형용사의 용법을 모두 가지고 있어 형태 표지 부여 방법을 명시할 필요가 있다.

1) 동사(VV)

동사는 사물의 움직임이나 작용을 나타내는 용언을 말한다. 동사는 일반적으로 목적어의 필요성 여부에 따라 자동사, 타동사로 나누기도 하지만, 본 분석에서는 그것을 위한 별도의 표지를 세분하지 않고 모두 'VV'로 표시한다.

2) 형용사(VA)

형용사는 사물의 성질이나 상태를 나타내는 용언을 가리킨다.

③ 관계언

관계언 지침에서는 조사의 결합형 분석과 관련하여 아래와 같이 조사를 분리해서 분석하도록 지시하고 있다. 하지만 '부사격조사' 지침에서는 '(으)로부터', '에서부터'를 하나

의 조사로 다루도록 하였다. 따라서 조사의 결합형 분석과 관련한 기준이 명시될 필요가 있다.

마. 관계언
조사는 주로 체언과 결합하여 다른 말과의 문법적 관계를 나타내거나, 특별한 뜻을 더해 주는 품사를 말한다. 조사는 크게 격조사, 보조사, 접속조사로 나눈다. 한국어는 조사가 중첩하는 경우가 많은데, 이러한 경우 조사의 결합형은 분리해서 분석함을 원칙으로 한다. 예: 부산에서도 대형 사고가 있었다. [부산/NNP+에서/JKB+도/JX] 그녀와의 약속이 갑자기 잡혔다. [그녀/NP+와/JKB+의/JKG]

④ 체언 접두사(XPN), 명사파생접미사(XSN)

아래 지침에서는 분리하여 분석할 접두사, 접미사의 목록을 제시하고 있다. 그런데 여기에도 몇 가지 불명확한 점이 포함되어 있다.

먼저 2음절 한자어에 아래 접두사, 접미사와 의미가 동일한 요소가 포함되어 있을 때 그것을 분리하여 분석할 것인지에 대한 지시가 포함되어 있지 않다. 가령 '최고(最高)'라는 단어를 '최+고'로 분리하여 분석해야 하는지 아닌지에 대한 명확한 언급이 없다.

또한 접사를 분리하는 데 있어서 단어의 내부 구조를 고려할 것인지 아닌지에 대한 지침이 포함되어 있지 않다. 가령 '부동산(不動産)'은 <우리말샘>에서 '부동-산'으로 제시되고 있는데, '부(不)'가 분석해야 하는 접두사 목록에 포함되어 있고 '동산(動産)'이 명사로 존재한다. 이 경우 '부동산'을 '부+동산'으로 분리하여 분석할 것인지, '부동산'의 직접 구성 요소(부동+산)를 고려하여 '부'를 분리하지 않을 것인지를 결정할 수 있도록 지침을 제시해야 한다.

가) 체언 접두사(XPN)

명사 접두사에는 한자어계 접두사와 고유어계 접두사가 있는데, 그 목록의 풍부함에 비해 대개가 생산성이 그리 높지 않다. 일단 여기서는 비교적 생산성이 높다고 인정되는 접두사와, 접두사를 분리했을 경우 단일한 표제어로 등재될 수 있는 경우에 한해서 접두사 분석을 하기로 한다.

가(假) 가건물
고(高) 고물가
과(過) 과보호
구(舊) 구소련
날 날음식

(후략)

가) 명사파생접미사(XSN)

명사파생접미사는 명사나 다른 어근에 후행하여 그것이 명사의 기능을 수행할 수 있도록 만들어 주는 의존 형태이다. 그러나 명사파생접미사는 연구자에 따라 그 목록이 다르며, 실제로도 구분이 애매한 경우가 많다. 본 분석에서는 접미사의 생산성과 접미사를 제외한 형태의 독립성을 기준으로 다음과 같이 목록을 마련하였다.

가(價) 매매가
가(哥) 김가
경(頃) 두 시경
계(系) 몽고계
계(界) 교육계

(후략)

1.1.2. 형태 표지 부여 방식이 재고되어야 하는 경우

① 일반명사(NNG)

<21세기 세종계획>의 형태 분석 말뭉치 구축 지침에서는 아래에서 볼 수 있듯이 '아이 러브 유'의 '아이', '벙커C유'의 '유'와 같이 직관적으로 일반명사로 보기 어려운 요소를 일반명사로 처리하도록 하였다. 이러한 요소들에 일반명사 표지를 부여하는 것이 적절한지 재고할 필요가 있다.

(1) 일반명사로 분석할 수 있는 단어	
(라) 외국어를 음차한 경우	
예: 아이 러브 유	[아이/NNG]
(마) 기타 다른 품사로 분석될 수 없는 단위	
예: 5관왕	[5/SN+관/NNG+왕/NNG]
벙커C유	[벙커/NNG+C/SL+유/NNG]

'아이 러브 유'는 외국어 문장을 음차한 것으로서 이 문장을 구성하고 있는 각 단어는 한국어에서 유의미하게 사용되는 언어 단위라고 보기 어렵다. 그럼에도 특정 품사를 부여하기 어려운 '아이'와 같은 요소에 일반명사 표지를 부여하도록 하였다.

또한 '벙커C유'는 전체를 한 단어로 볼 수 있음에도 단어 가운데 'C'라는 로마자가 포함되어 있기 때문에 '벙커'와 'C', '유'를 분리하여 분석하도록 하였고, '유'는 단독으로 쓰이지 않아 명사로 보기 어려운 요소임에도 일반명사로 처리하도록 하였다.

이러한 문제를 해결하기 위해 외국어 요소를 어떤 방식으로 처리하는 것이 우리말의 구성 단위를 유의미하게 보여 줄 수 있는 방식인지 고민할 필요가 있다. 또한 단어 자격을 갖지 않는 요소를 과도하게 단어로 분석하지 않도록 할 방법을 강구할 필요가 있다.

② 고유명사(NNP)

<21세기 세종계획>의 형태 분석 말뭉치 구축 지침에서는 '고유명사의 분석 기준은 매우 다양하므로, 본 지침에서는 다음에 제시하는 것만을 고유명사로 인정한다.'라고 하면서 고유명사로 인정하는 부류로 '인명・종족명, 지명, 국가명・왕조명, 건축물・시설물・구조물, 회사・학교・정당・기관・단체, 창작물의 제목, 언어명'을 제시하였다.

이에 따르면 배, 비행기와 같은 건조물의 이름(예: 최영함), 상품명(예: 코카콜라), 브랜드명(예: 래미안), 상호명(예: 한마음약국), 팀명(예: 트와이스), 매체명(예: 여성동아, 트위터, 네이버) 등이 고유명사 부류에 포함되지 않으며, 이들 부류를 고유명사에서 제하고 일반명사로 취급하는 것은 언중의 직관에 반한다는 점에서 문제가 있다.

따라서 언중의 직관을 고려하면서도 작업의 일관성을 위해 고유명사로 처리할 부류의 기준을 분명히 할 수 있는 방안을 강구할 필요가 있다.

③ 종결어미(EF), 연결어미(EC)

<21세기 세종계획>의 형태 분석 말뭉치 구축 지침에서는 종결어미, 연결어미의 구분과 관련하여 아래와 같은 지침을 제시하였다. 즉 문장 부호를 종결어미와 연결어미를 구분하는 기준으로 삼아, 마침표, 물음표, 느낌표 등 종결 부호(SF) 앞에서 사용된 어미는 종결어미로, 그렇지 않은 어미는 연결어미로 분석하도록 한 것이다.

나) 종결 어미(EF)
용언의 어간이나 선어말 어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 한 문장을 끝맺는 역할을 하는 어미이다. 그러나 종결어미가 문장의 종결에만 사용되는 것은 아니고, 문맥에 따라서는 연결 어미로 쓰이기도 한다. 본 지침에서는 "SF" 앞에서만 종결어미를 인정한다.

종결어미와 연결어미의 구분이 모호한 경우가 많으므로 이러한 지침을 적용하면 결과물의 신뢰성(일관성)을 도모할 수 있다는 장점이 있지만, 이는 문법적 정확성보다 작업의 편리성을 추구한 지침에 해당하며 분석의 타당성을 도모하기는 어렵다.

가령 이 지침에 따르면 신문 기사 제목에서 '골수 기증 열기 휴일도 없다'와 같이 문장 부호를 사용하지 않았을 경우 문말의 '다'를 연결어미로 분석하게 되어 타당하지 않은 분석 결과를 낳게 된다. 문장 부호가 있느냐 없느냐는 종결어미와 연결어미 구분의 결정적인 기준은 아니므로, 좀 더 언어학적 타당성을 갖춘 기준에 따라 판단할 필요가 있다.

④ 기호

기호가 포함된 어절의 처리에 대해서는 아래와 같은 지침을 제시하였다. 어절에 로마자나 한자, 기호 등이 포함된 경우에는 그 앞뒤 요소를 모두 분리하여 분석하도록 한 것이다.

1) 기호

영문이나 한자, 기호 등이 어절 중간에 개입하여 올바른 분석이 불가능한 경우에는 각각의 요소를 분리하여 분석한다. 이 경우 표지를 줄 수 없는 불완전한 형태가 생길 수 있다.

예: 마이크로소프트(microsoft)사	[마이크로소프트/NNP+(/SS+microsoft/SL+)/SS+사/NNG]
농·수산물	[농/NNG+·/SP+수산물/NNG]
초·중·고	[초/NNG+·/SP+중/NNG+·/SP+고/NNG]

그러나 이처럼 분석하면 '농・수산물'의 '농'과 같이 품사를 부여하기 어려운 요소가

남게 되는 경우가 많다. 또 다른 예로 '4인승'은 <21세기 세종계획>의 지침에 따르면 '4/SN+인/NNG+승/NNG'으로 분석되는데, 이때 '승'은 국어에서 명사 자격을 가지며 홀로 쓰이는 말이라고 보기 어려우므로 언중이 단어로 인식하지 않는 단위를 과도하게 분석한다는 문제를 안게 된다.

따라서 단어 단위에 대한 언중의 직관을 고려하여 분석 수준을 재고할 필요가 있다.

1.2. 지침의 보완 방향

<21세기 세종계획> 형태 분석 말뭉치 구축 지침을 검토한 결과, 아래와 같은 보완의 필요성이 제기되었다.

○ 언중의 직관을 고려하여 고유명사의 범위를 보다 넓히면서도, 작업의 일관성을 위하여 구체적으로 범위를 명시하고 다양한 사례를 제시할 필요가 있다.
○ '있다'의 동사 용법과 형용사 용법을 구분하는 기준을 제시할 필요가 있다.
○ 조사 결합형을 분석하는 기준을 명시할 필요가 있다.
○ 접두사, 접미사를 언제 분리하는지에 대한 지침을 명시할 필요가 있다.
○ 한국어에서 유의미하게 사용되는 언어 단위를 분석 대상으로 삼기 위하여 외국어 처리 지침과 기타 기호가 포함된 어절의 처리 지침을 재정비할 필요가 있다.
○ 종결어미와 연결어미를 후행하는 문장 부호에 따라 구분하기보다 기능에 따라 구분하도록 해야 한다.

위의 보완점들은 아래와 같이 '분석 지침의 구체화, 언어학적 엄밀성의 추구, 과도한 분석 지양'의 세 유형으로 재분류될 수 있다. 이 세 가지는 본 사업에서 기존 지침을 수정・보완할 때 추구한 방향성에 해당한다.

○ 분석 지침의 구체화

- 조사 결합형을 분석하는 기준을 명시할 필요가 있다.
- 접두사, 접미사를 언제 분리하는지에 대한 지침을 명시할 필요가 있다.
- 언중의 직관을 고려하여 고유명사의 범위를 보다 넓히면서도, 작업의 일관성을 위하여 구체적으로 범위를 명시하고 다양한 사례를 제시할 필요가 있다.

○ 언어학적 엄밀성의 추구

- '있다'의 동사 용법과 형용사 용법을 구분하는 기준을 제시할 필요가 있다.
- 종결어미와 연결어미를 후행하는 문장 부호에 따라 구분하기보다 기능에 따라 구분하도록 해야 한다.

○ 과도한 분석 지양

- 한국어에서 유의미하게 사용되는 언어 단위를 분석 대상으로 삼기 위하여 외국어 처리 지침과 기타 기호가 포함된 어절의 처리 지침을 재정비할 필요가 있다.

1.2.1. 분석 지침의 구체화

① 조사 결합형의 분석 기준 명시

조사 결합형 중에는 '에는'처럼 두 조사의 의미가 투명하게 드러나고 조사를 분리하여 사용하여도 문제가 없어 단순히 조사와 조사가 연쇄된 것이 있는가 하면, '로부터'처럼 두 조사가 결합하여 출발점이라는 새로운 의미를 나타내고 조사를 분리하여 사용하면 문장이 성립하지 않아 하나의 조사가 된 것으로 볼 수 있는 것이 있다. 전자의 조사 결합형은 그것을 이루는 조사들을 각각 분리하여 형태 표지를 부여하는 것이 바람직하고, 후자의 조사 결합형은 그 내부를 더 분리하지 않고 조사 결합형 전체에 하나의 형태 표지를 부여하는 것이 바람직하다.

하지만 조사 결합형의 두 유형을 엄밀하게 구분하여 목록화하는 일은 쉽지 않다. 이에

본 사업에서는 조사 결합형에 대한 <우리말샘>의 뜻풀이 방식을 참고로 삼아 분리하여 분석할 조사 결합형과 분리하지 않을 조사 결합형을 구분하기로 하였다. 가령 위에서 예로 든 '에는'과 '로부터'는 <우리말샘>에서 다음과 같이 뜻풀이되고 있다.

(1) ㄱ. 에-는: 「조사」 부사격 조사 '에'에 보조사 '는'이 결합한 말. 강조와 대조의 뜻을 나타내는 조사이다.

ㄴ. 로-부터: 「조사」((받침 없는 체언이나 'ㄹ' 받침으로 끝나는 체언 뒤에 붙어)) 어떤 행동의 출발점이나 비롯되는 대상임을 나타내는 격 조사. 격 조사 '로'와 보조사 '부터'가 결합한 말이다.

위에서, 단순한 조사 연쇄형인 '에는'은 '~이 결합한 말'의 형식으로, 하나의 조사가 된 것으로 볼 수 있는 '로부터'는 '~을 나타내는 격 조사'의 형식으로 뜻풀이된 것을 확인할 수 있다.

물론 이와 같은 <우리말샘>의 서로 다른 뜻풀이 방식이 엄밀한 기준에 따라 이루어진 것인지에 대해서는 향후 지속적인 검토가 이루어져야 하지만, 조사 결합형의 두 유형을 나누면서도 일관성 있는 처리를 하기 위해서는 <우리말샘>의 뜻풀이 방식을 기준으로 삼아 조사 결합형의 조사를 분리할 것인지 통합할 것인지를 결정하는 것이 현실적 방안이 될 수 있다고 판단하였다. 이에 아래와 같이 조사 결합형의 처리 원칙을 명시하였다.

주의사항

한국어는 조사가 여러 개 결합하는 경우가 많은데, 조사 결합형은 아래와 같은 방식으로 세분 여부를 결정한다.

① 조사 결합형이 <우리말샘>에 등재되어 있지 않으면 각 조사를 분리하여 분석한다.

[예시] 부산에서도 대형 사고가 있었다. [부산/NNP+에서/JKB+도/JX] (‘에서도’ 미등재)
그녀와의 약속이 갑자기 잡혔다. [그녀/NP+와/JKB+의/JKG] (‘와의’ 미등재)

② 조사 결합형이 <우리말샘>에 등재되어 있으면, 사전의 뜻풀이를 참고하여 결합형 자체에 ‘격 조사’나 ‘보조사’라고 풀이되어 있으면 더 분석하지 않고 하나의 조사로 둔다.

[예시] 에다가 [에다가/JKB]
(사전: 일정한 위치를 나타내는 격 조사. 격 조사 ‘에’에 보조사 ‘다가’가 결합한 말이다.)

③ 만약 조사 결합형이 <우리말샘>에 등재되어 있는데 ‘어떤 조사와 어떤 조사가 결합한 말’로만 풀이되어 있으면 두 개의 조사로 분리하여 분석한다.

[예시] 에는 [에/JKB+는/JX]
(사전: 부사격 조사 ‘에’에 보조사 ‘는’이 결합한 말. 강조와 대조의 뜻을 나타내는 조사이다.)

② 접두사, 접미사의 분리 기준 명시

접두사, 접미사의 분리와 관련하여 우선 고려해야 할 것은 한자어 접두사, 접미사의 분리 문제와 관련된 것이다. 가령 한자어 접미사 ‘-가(價)’는 3자어인 ‘소매가(小賣價)’에서는 접미사로 인식되지만 2자어인 ‘유가(油價)’에서는 접미사로서의 인식이 약하다. 이처럼 국어 화자들은 한자어에서 의미를 가진 음절 하나하나를 인식할 수 있지만, 2자어의 구성 요소인 1자들이 가지는 자격은 순전히 의미적인 것일 뿐, 품사적 자격을 비롯해서 어떤 문법 기능을 나타낼 수 있는 단위로는 여겨지지 않는다[1]. 따라서 한자어 2자어에 분석 대상이 되는 접두사, 접미사와 형식·의미가 동일한 요소가 들어 있을 때에는 그 요소가 접두사, 접미사로서의 자격을 가지고 사용된 것이 아니라고 볼 수 있으며, 한

1) 김창섭(2001), 한자어 형성과 고유어 문법의 제약, 『국어학』 37, 국어학회.

자어 접두사, 접미사는 3자 이상의 어휘에서 분리하는 것이 타당하다.

이러한 인식은 <우리말샘>의 표제어 형식에도 반영되어 있다. '소매가'는 '소매-가'로 표제어에 하이픈이 포함되어 있지만, '유가'는 하이픈 없이 '유가'로 등재되어 있는 것이다. 이를 고려하여 본 사업의 지침에서는 <우리말샘>의 표제어에 하이픈이 있는 경우에만 접두사, 접미사를 분리하도록 하였다.

또한 접두사, 접미사 분리 시 '어근' 단위가 남는 경우에도 접두사, 접미사를 분리할 것인지를 결정해야 한다. <21세기 세종계획>의 형태 분석 말뭉치 구축 지침과 <우리말샘>에서 다루는 '어근'의 개념은 범위의 차이가 있을 수 있으나 대체로 자립성이 없으며 제한된 요소와 결합해서만 쓰이는 요소를 말한다. 어근은 일반 언중에게 독자적인 언어 단위로서 분명히 인식되지 않는 경우가 많으므로, 접사를 분리했을 때 어근이 남는다면 접사 분리를 하지 않는 것이 접사의 생산성을 잘 보여 줄 수 있는 방법일 뿐 아니라 단어 단위를 중심으로 형태 분석을 하는 본 사업의 태도에도 부합한다. 이에 접사를 분리할 경우 어근이 남는다면 접사 분리를 하지 않도록 명시하였다.

마지막으로 접사 분리 시 단어의 직접 구성 요소를 고려할 수 있도록 지침을 제시할 필요가 있다. 단어의 직접 구성 요소에 대한 정보는 <우리말샘>의 하이픈 위치에서 얻을 수 있으므로, <우리말샘>에서 단어를 검색하여 하이픈의 바로 앞에 분석 대상 접두사가 있거나 하이픈의 바로 뒤에 분석 대상 접미사가 있는 경우에만 접사를 분리해 내도록 하였다. 이 지침을 따르면 '부동-산(不動産)'에는 '부(不)'라는 요소가 들어 있지만 이 단어의 직접 구성 요소가 아니기 때문에 접두사 분리를 하지 않고 '부동산/NNG'로 형태 표지를 부여하게 된다. '비합리-적(非合理的)'처럼 한 단어에 접사가 두 개 들어 있는 경우도 있는데, 이때에도 직접 구성 요소를 고려하여 '비합리+적'으로 먼저 분리하고, 남은 단어인 '비-합리(非合理)'를 한 번 더 '비+합리'로 분리함으로써 '비/XPN+합리/NNG+적/XSN'으로 분석하도록 구체적인 지침을 마련하였다.

위와 같은 점을 고려하여 마련한 접사 분리 지침은 아래와 같다.

주의사항

접사(접두사, 접미사)는 아래 가)~라)에 목록화된 접사가 등장한 경우에만 분리하여 분석한다. 접사의 분리 원칙은 다음과 같다.

2음절 단어의 처리

① <우리말샘> 등재어인 경우(즉 원어절에 나타난 2음절 단어와 같은 의미의 단어가 <우리말샘> 표제어로 올라 있는 경우), 표제어에 하이픈이 있는 경우에만 접사를 분리한다. 단, 접사를 분리하고 남는 요소가 어근에 해당하는 경우에는 접사를 분리하지 않는다.

[예시] 오형 (사전: 오-형, 혈액형의 하나) [오/NNG+형/XSN]

<우리말샘>에 명사, 부사로 등재된 의미나 쓰임이 아니라, 그 앞뒤로 다른 말(단위성 의존명사 등)과 함께 쓰이는 '순서'의 '제일'은 제/XPN+일/NR로 분리한다.

[예시] 제일 차(회/조/항…) 회의 [제/XPN+일/NR] (○)
[제일/NNG] (×), [제일/MAG] (×)

분석 대상이 되는 말이 <우리말샘>에 등재되어 있다고 하더라도 그 쓰임과 의미를 면밀히 확인할 수 있도록 주의한다.

[예시] 15일자 신문 [15/SN+일/NNB+자/NNG] (○)
[15/SN+일자/NNG] (×)

② <우리말샘> 미등재어인 경우, 그 단어가 2음절 한자어이거나 접사 분리 시 어근이 남는다면 접사를 분리하지 않는다. 그 외의 경우에는 접사를 분리한다.

[예시] 뇌성마비(사전: 뇌성^마비) [뇌성/NNG+마비/NNG]
→ <우리말샘>에서 '뇌성'은 단독 표제어로 올라 있지 않아 하이픈 유무를 참고할 수 없다. 하지만 2음절 한자어에 해당하는 말이 <우리말샘>에서 대체로 하이픈 없이 처리되고 있음을 참고하여 '-성'을 분리하지 않고 명사로 처리한다.

[예시] 나는 아침형 인간이 아니라 밤형 인간이다. [밤/NNG+형/XSN]
→ '밤형'은 미등재어이지만 2음절 한자어가 아니다. 또한 '밤'이 명사이므로 '밤'과 '-형'을 분리한다.

③ 숫자, 로마자 등 기타기호에 접사가 결합한 것은 일반명사 지침의 (라)항을 우선 적용하여 처리한다.

[예시] 3분의 일 [3/SN+분/XSN]
→ 명사 '삼분'이 <우리말샘>에 하이픈 없이 등재되어 있지만, 기타기호의 처리 방법을 우선 적용하여 [3/SN+분/XSN]으로 분리하여 분석한다.

3음절 이상 단어의 처리

① 3음절 이상 복합어의 경우, <우리말샘>의 표제어 하이픈(-) 위치를 참고하여 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있는 경우에만 해당 접사를 분리한다.

[예시] 과보호(사전: 과-보호) [과/XPN+보호/NNG]
→ 하이픈 바로 앞에 놓인 접사인 '과-'가 분석 대상 접사이다. 이 경우 '과'와 '보호'를 분리한다.
[예시] 피보험자(사전: 피보험-자) [피보험자/NNG]
→ 하이픈 바로 뒤에 놓인 접사인 '-자'는 본 지침의 분석 대상 접사가 아니다. 따라서 더 이상 분리하지 않고 전체를 [피보험자/NNG]로 분석한다.

② 만약 접사를 분리했을 때 남는 단위가 어근(XR)이라면 접사 분리를 하지 않는다.

[예시] 비롯하다(사전: 비롯-하다) [비롯하/VV+다/EF]
→ 하이픈 바로 뒤에 놓인 접사인 '-하-'가 분석 대상 접사이지만, 이것을 분리하고 남는 단위인 '비롯'이 어근에 해당한다. 이 경우 '비롯'과 '하'를 분리하지 않는다.

③ 접사를 분리하고 남은 부분이 사전 미등재어인 경우가 있다. 그 미등재어가 홀로 쓰이지 않아 어근 자격을 갖는 것으로 판단된다면, 위 ②와 마찬가지로 접사를 분리하지 않는다.

[예시] 역세권(사전: 역세-권) [역세권/NNG]
→ 하이픈 바로 뒤에 놓인 접사인 '-권'이 분석 대상 접사이지만, 이것을 분리하고 남는 단위인 '역세'가 사전 미등재어이며 홀로 쓰이지도 않아 어근에 해당한다. 이 경우 '역세'와 '권'을 분리하지 않는다.

접사를 분리하고 남은 미등재어가 합성어라면, 합성어 분석 원칙에 따라 합성어 구성 요소를 분리하여 분석한다.

[예시] 중고생(사전: 중고-생) [중/NNG+고/NNG+생/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-생'이 분석 대상 접사인데, 이것을 분리하고 남는 단위인 '중고'가 사전 미등재어이다. 그런데 사전에 중학교를 뜻하는 '중', 고등학교를 뜻하는 '고'가 명사로 올라 있어 이 말은 합성어로 파악된다. 이 경우 접사를 분리하고 남은 미등재 합성어를, 여타 미등재 합성어의 처리 방식과 마찬가지로 분리하여 분석한다.

[참고] 일회용(사전: 일회-용) [일회/NNG+용/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-용'이 분석 대상 접사인데, 이것을 분리하고 남는 단위인 '일회'가 단독으로는 사전 등재어가 아니다. 하지만 '일회^결실성' 등의 구 표제어 속에서 한 단어로 나타나므로 더 분석하지 않고 한 단어로 취급한다.

④ 만약 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있어서 해당 접사를 분리해 냈는데, 접사를 뗀 나머지 부분에 또 분석 대상 접사가 포함되어 있을 수 있다. 그런 경우에는 그 나머지 단어를 <우리말샘>에서 검색하여 하이픈의 위치를 확인한 후, 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있다면 해당 접사를 다시금 분리해 낸다.

[예시] 비합리적(사전: 비합리-적) [비/XPN+합리/NNG+적/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-적'이 분석 대상 접사이므로 '비합리'와 '적'을 분리한다. 그런데 '적'을 뗀 나머지 부분인 '비합리'(사전: 비-합리)에 분석 대상 접사인 '비'가 들어 있고, <우리말샘>에서 '비합리'를 검색했을 때 하이픈 바로 앞에 '비'가 놓여 있다. 이 경우 '비'와 '합리'를 다시금 분리한다. 결과적으로 '비합리적'을 [비/XPN+합리/NNG+적/XSN]으로 분석하게 된다.

⑤ 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있어서 해당 접사를 분리했을 때 남는 단위가 2음절 요소라면 위 '2음절 단어의 처리'에 따라 해당 2음절 요소를 처리한다.

⑥ 복합어가 <우리말샘> 미등재어여서 하이픈 정보를 참고할 수 없는 경우에는 <우리말샘> 등재 어휘를 참조하고 복합어의 의미 구조에 대해 직접 판단하여 처리한다.

[예시] 최대형 (미등재어) [최대/NNG+형/XSN]
→ 사전 등재어인 '최소형'(사전: 최소-형)을 참고하여 처리할 수 있다.
[예시] 대의원회 (미등재어) [대의원회/NNG]
→ '대의원의 모임'이라는 뜻이므로 의미 구조상 '대의원-회'로 나뉜다. 이때 하이픈 뒤의 '-회'는 본 지침의 분석 대상 접사가 아니므로 이 단어를 더 분리하지 않는다.

③ 고유명사 지침 구체화

앞서 살펴보았듯이 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침에서는 고유명사의 범위를 좁게 설정한 바 있다. 이는 고유명사와 일반명사를 구분하는 일이 본디 쉽지 않고 따라서 고유명사 판별 과정에 작업자들의 직관 차이가 작용함으로써 최종 결과물의 일관성을 저해할 수 있다는 판단 때문인 것으로 생각된다.

하지만 고유명사의 범위를 좁게 설정하다 보니 배, 비행기와 같은 건조물의 이름(예: 최영함), 상품명(예: 코카콜라), 브랜드명(예: 래미안), 상호명(예: 한마음약국), 팀명(예: 트와이스), 매체명(예: 여성동아, 트위터, 네이버) 등 언중이 분명히 고유명사로 인식하는 요소가 고유명사 부류에 들지 못해 반직관적이라는 문제가 있었다. 특히 통신 수단의 발달로 트위터, 유튜브, 개인 채널 등 이전과는 다른 매체가 많이 등장하게 되었으므로 시대의 변화도 반영하여 고유명사의 범위를 재정비할 필요가 있다.

본 사업에서는 언중의 직관을 고려하면서도 작업의 일관성을 위해 고유명사로 처리할 부류의 기준을 분명히 할 수 있는 방안을 강구하고자 하였다. 이에 건조물의 이름, 상품명과 브랜드명, 상호명, 팀명, 매체명을 고유명사 부류로 추가하였다. 그리고 고유명사로 처리할 부류의 기준을 분명히 하기 위해 해당 부류에 속하는 예를 다양하게 제시하였다.

한편 기존 지침의 '지명' 부분에 예가 충분히 포함되지 않아 '관동팔경, 경인공업지대, 가자지구, 경포해수욕장, 광화문광장' 등 일반적으로 지명으로 인식되는 명칭을 고유명사

로 처리할 것인지 아닌지를 지침만으로는 판단하기 어려웠음을 앞서 언급한 바 있다. '광화문광장'은 시설물이나 구조물로 볼 수도 있는데, '건축물이나 시설물 혹은 구조물의 이름' 지침에서도 이러한 사례를 포함하고 있지 않았다. 이를 보완하기 위해 '지명'과 '건축물, 시설물, 구조물' 지침에 보다 다양한 종류와 사례를 추가하였다.

단체명을 어디까지 고유명사로 처리할 것인지도 좀 더 구체화할 필요가 있었다. 이에 회사, 학교, 학회, 협회, 정당의 이름은 모두 고유명사로 분석하도록 명시하였고, 정부기관의 명칭 중 고유명사를 포함하고 있는 것만을 고유명사로 분석하도록 하되 특정 기관의 '지청, 지원, 지부' 등은 그것이 지명과 함께 나타났더라도 모든 기관에 존재할 수 있어 특정성이 낮으므로 고유명사로 처리하지 않도록 지침을 구체화하였다. 연구소, 위원회, 협의회, 본부, 종교단체는 고유명사나 '국가, 전국, 국제, 세계' 등을 포함하는 것을 고유명사로 처리하도록 하였고, 예시를 추가하였다. 또한 기존 지침에는 단체장명 분석 지침이 포함되어 있지 않아 '서울시장'은 '서울/NNP+시장/NNG'로, '중랑구청장'은 '중랑구청장/NNG'로 분석하는 등 단체장명의 분석 방식이 통일되지 않았었는데, 지명/단체명에 접사 '-장(長)'이 결합한 것으로 보아 '서울시장/NNG', '중랑구청장/NNG'로 분석하도록 명시하였다.

보완된 고유명사 지침 중 위에서 언급한 내용과 관련 있는 부분을 아래에 제시하였다.

나) 고유명사(NNP)

(2) 지명

<우리말샘>에 『지명』으로 올라 있는 단어 부류를 참고하여 지명 여부를 판단하고, 지명에 해당하는 부분과 지역의 종류를 나타내는 말을 묶어 전체를 고유명사로 처리한다.

(가) 내륙, 대륙, 지대, 주, 평원, 만, 늪, 습지, 분지, 사막, 유전, 탄전, 군락지
섬, 제도, 열도
바다, 해변, 포구, 강, 유역, 나루, 호수, 계곡, 연못, 갯벌, 폭포, 삼각주, 빙하, 피오르

길, 거리, 수로, 루트, 로드
산, 산맥, 산지, 화산, 동굴, 숲, 고개, 언덕, 오름, 구릉, 고원, 광산, 절리, 화산대 등의 이름

[예시] 카스피해/NNP, 템즈강/NNP, 태백산맥/NNP, 미시시피호/NNP, 갈라파고스제도/NNP, 갈론계곡/NNP, 감지해변/NNP, 강경포구/NNP, 강계분지/NNP, 강주연못/NNP, 거창분지/NNP, 고수동굴/NNP, 관동팔경/NNP, 그레이트빅토리아사막/NNP

(나) 도(道), 시(市), 읍(邑), 면(面), 리(里), 군(郡), 구(區), 동(洞), 골, 촌, 마을, 자치구, 연구단지, 관광단지, 공업지대, 지역, 지방, 지구 등의 이름은 그 구역의 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 서울특별시/NNP, 성북구/NNP, 강진군/NNP, 인창동/NNP, 빨래골/NNP, 해방촌/NNP, 고포마을/NNP, 네바다주/NNP, 경기지역/NNP, 경인공업지대/NNP, 관서공업지역/NNP, 광시좡족자치구/NNP, 가자지구/NNP, 서안지구/NNP

(4) 건축물이나 시설물 혹은 구조물의 이름

<우리말샘>에 『지명』으로 올라 있는 단어 부류를 참고하여 고유명사 여부를 판단하고, 구조물명, 시설물명에 해당하는 부분과 구조물, 시설물의 종류를 나타내는 말을 묶어 전체를 고유명사로 처리한다.

(가) 도로, 항만, 항구, 터널, 대교, 철교, 뱃길, 운하, 댐, 공항, 터미널, 철도, 전철, 지하철 및 그 명칭과 함께 쓰이는 부대시설은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 부산항/NNP, 대전역/NNP, 서울지하철/NNP, 인천공항/NNP, 테헤란로/NNP, 경부고속도로/NNP, 분당선/NNP, 경춘선/NNP, 정우터널/NNP, 강화대교/NNP, 경인아라뱃길/NNP

단, 어느 지역의 지하철에나 존재하는 '1호선, 2호선' 등은 특정성이 낮으므로 고유명사로 보지 않는다.

[예시] 1호선 [1/SN+호선/NNB]
→ 의존명사 '호'에 비분석 접사 '-선'이 결합한 구성으로, '-선'을 앞말에 붙여 '호선/NNB'로 처리한다.

(나) 해수욕장, 공원, 광장, 정원, 목장, 유원지, 유적지, 절터, 관광지, 테마파크, 전망대, 온천, 시장, 장터, 저수지, 기지, 묘지 등의 시설물도 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 속초해수욕장/NNP, 올림픽공원/NNP, 설악산국립공원/NNP, 한강시민공원/NNP, 광화문광장/NNP, 화정역광장/NNP, 황복사터/NNP, 가락시장/NNP, 노량진수산시장/NNP, 남극세종기지/NNP, 도라전망대/NNP, 현충원/NNP, 국립서울현충원/NNP

단, '농산물도매시장', '생활체육공원'과 같이 특정 시장이나 공원의 이름이 아니라 시장이나 공원의 유형을 나타내기 위해 쓰인 말은 고유명사로 보지 않는다.

'문화예술공원'은 서울특별시 서초구에 있는 특정 공원의 이름으로 쓰이기도 하고 공원의 유형을 나타내기 위한 말로 쓰이기도 하는데, 맥락을 구분하여 전자의 경우에는 고유명사로, 후자의 경우에는 일반명사로 분석한다.

[예시] 서초구 문화예술공원 (특정 공원의 이름일 때)
[문화예술공원/NNP]
우리 구에 문화예술공원을 설립합시다. (공원의 종류를 나타낼 때)
[문화/NNG+예술/NNG+공원/NNG]

(다) 배, 비행기와 같은 건조물의 이름은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다. [참고]와 같이 특정 집단의 우두머리 이름을 따 배에 빗대어 표현하는 경우에도 고유명사로 처리한다.

[예시] 최영함/NNP, 나로호/NNP
[참고] 신태용호/NNP

(라) 빌딩, 박물관, 극장 등 건물명은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다. 그리고 <우리말샘>에 구로 등재된 단위라고 하여도 그 통합형을 고유명사로 처리한다.

[예시] 서울역사/NNP, 세종문화회관/NNP, 개나리유치원/NNP, 청와대/NNP, 국회의사당/NNP
국립중앙박물관/NNP, 국립민속박물관/NNP, 구텐베르크박물관/NNP
신라호텔/NNP, 미도파백화점/NNP, 동궁예식장/NNP, 명보극장/NNP, 고대병원/NNP, 인천타워/NNP, 고마신사/NNP

(5) 회사, 상품, 학교, 정당, 기관이나 단체의 이름

(가) 특정 회사, 학교, 학회, 협회, 정당의 이름은 고유명사로 분석한다.

[예시] 삼성/NNP, 삼성그룹/NNP, 엘지전자/NNP, LG전자/NNP, 현대자동차서비스/NNP
코스닥/NNP, 나스닥/NNP, 코스피/NNP
고려대학교/NNP, 잠실고등학교/NNP, 송파중학교/NNP
국어학회/NNP, 영어영문학회/NNP, 한국사학회/NNP
대한축구협회/NNP, 승마협회/NNP, 프로야구선수협회/NNP
한나라당/NNP, 자유민주주의연합/NNP

회사, 학교, 학회, 협회, 정당의 이름이라고 해도 아래와 같이 한글 없이 기타 문자로만 표기된 경우에는 고유명사가 아니라 기호로 처리한다. 인명 등도 마찬가지이다.

[예시] LG에서 [LG/SL+에서/JKB]

(나) 특정 회사의 상품명과 브랜드명은 전체를 묶어 고유명사로 처리한다. 회사명과 상품명이 한 어절에 나왔을 때나 상품명과 상품의 종류를 나타내는 말이 한 어절에 나왔을 때에도 전체를 묶어 고유명사로 처리한다.

[예시] 초코하임/NNP, 코카콜라/NNP, 갤럭시S8/NNP, e편한세상/NNP, 쉐보레/NNP, 티맵/NNP
[예시] 농심새우깡/NNP, 칠성사이다/NNP, 신한SOL/NNP (회사명+상품명)
[예시] e편한세상아파트/NNP (브랜드명+종류명)
[예시] 하나에스라인적금/NNP (회사명+상품명+종류명)

(다) 상호명은 그 통합형을 고유명사로 처리한다. 상호명과 업종명이 한 어절에 나타난 경우에는 전체를 통합하여 고유명사로 분석하고, 상호명과 업종명이 두 어절로 분리되어

나타난 경우에는 상호명에 해당하는 부분을 통합하여 고유명사로 분석한다.

[예시] 한마음약국/NNP, 동일카센터/NNP
[예시] 다나은 이비인후과 [다나은/NNP, 이비인후과/NNG]

(라) 정부기관의 명칭 중 **지명 등의 고유명사를 포함하고 있어** 특정성이 높은 것만을 통합하여 고유명사로 처리한다. <우리말샘>에 구로 등재된 단위라고 하여도 그 통합형을 고유명사로 처리한다.

단, 특정 기관의 '지청, 지원, 지부', 특정 정당 소속의 '시당' 등은 그것이 지명과 함께 나타났더라도 모든 기관에 존재할 수 있어 특정성이 낮으므로 고유명사로 처리하지 않고 [예시]와 같이 단어별로 분리하여 일반명사로 분석한다.

하지만 '지방법원'의 준말인 '지법'은 지명과 함께 쓰일 경우 특정 기관을 가리키므로(예: 서울지법, 부산지법), '서울지법/NNP, 부산지법/NNP'로 처리한다.

고유명사를 포함하지 않아 고유명사에 들지 않는 정부기관 명칭은 단어별로 분리하여 처리한다.

[예시] 서울고등법원/NNP, 서울시경찰서/NNP, 서대문구치소/NNP
[예시] 여주/NNP+지청/NNG, 충북/NNP+지부/NNG, 서울시/NNP+당/NNG(서울시+당[명사])
[예시] 헌법/NNG+재판소/NNG, 국립/NNG+국어원/NNG, 여성/NNG+가족부/NNG

어떤 단어가 정부기관의 명칭인지, 건물명인지 혼동되는 경우가 있다. 이때에는 <우리말샘>의 뜻풀이를 참고하여, 주된 뜻풀이에 '기관'이라 되어 있으면 정부기관으로 판단하고, '건물'이라 되어 있으면 건물명으로 판단한다.

[예시] 청와대 (사전: 서울 경복궁 뒤 북악산 기슭에 있는 우리나라 대통령 **관저**)
→ 사전의 뜻풀이에서 건물명으로 처리되고 있으므로 [청와대/NNP]로 처리한다.
[예시] 경찰청 (사전: 안전 행정부 소속하에 설치되어 경찰 업무를 관장하는 정부 행정 **기관**)
→ 사전의 뜻풀이에서 정부기관명으로 처리되고 있으며 고유명사가 포함되어 있지 않으므로 [경찰청/NNG]로 처리한다.

(마) 연구소, 위원회, 협의회, 본부, 종교단체의 경우에는 **고유명사나 '국가, 전국, 국제, 세계' 및 이와 유사한 단어를 포함하고 있어** 특정성이 높은 것만을 통합하여 고유명사로 처리한다.

고유명사에 들지 않는 연구소, 위원회, 협의회, 본부, 종교단체명은 단어별로 나누어 처리한다.

[예시] 한국전자통신연구소/NNP, 대한예수교장로회/NNP, 대한불교조계종/NNP, 서울자사고교장협의회/NNP, 한국야구위원회/NNP, 국가인권위원회/NNP, 국제통화기금/NNP
[예시] 조계종/NNP, 천태종/NNP (인명, 지명을 포함하고 있는 말임)
[참고] 감리교/NNG, 장로교/NNG (고유명사를 포함한 말이 아님)
[예시] 통일/NNG+연구소/NNG, 생활/NNG+체육/NNG+연구소/NNG,
통사/NNG+론/XSN+연구회/NNG, 입주자/NNG+대표/NNG+협의회/NNG,
수니파/NNG, 방송/NNG+통신/NNG+위원회/NNG
→ 이처럼 고유명사나 '국가, 전국, 국제, 세계'를 포함하지 않은 단체명은 단어별로 나누어 처리한다.

(바) 특정 부대, 스포츠팀, 그룹, 조직, 클럽에 붙여진 이름은 종류를 나타내는 말과 함께 묶어 고유명사로 처리한다. 이 밖에도 이 부류와 유사한 특정 집단에 고유하게 붙여진 이름을 고유명사로 처리한다.

[예시] 백마부대/NNP (부대명) cf) '육군', '해군', '공군', '해병대'는 일반명사임.
인디언스/NNP, 기아/NNP, 타이거즈/NNP, 서울FC/NNP (스포츠팀명)
트와이스/NNP, 송골매/NNP (그룹명)
서방파/NNP (폭력조직명)
위너스클럽/NNP (클럽명)
티파티/NNP (미국 정부의 건전한 재정 운용을 위해 세금 감시 운동을 펼치는 시민 중심의 신생 보수 단체.)

(아) 단체장명의 처리

지명/단체명에 '-장(長)'과 같이 접미사가 결합하거나 '수(守)'와 같이 접사에 준하는 요소가 결합하여 단체장명이 만들어진 경우, 해당 요소를 바로 앞말에 붙여 분석한다. (가령 '헌법재판소'는 본 지침에서 '헌법/NNG+재판소/NNG'가 되므로, 여기에 '-장'이 결합한 경우 '헌법/NNG+재판소장/NNG'으로 처리한다.) 또한 단체장명은 일반명사임에 유의하여 형태 표지를 부여한다.

[예시] 가평군수(가평군/NNP+수) [가평군수/NNG]
서울시장(서울시/NNP+장) [서울시장/NNG]

헌법재판소장(헌법/NNG+재판소/NNG+장)	[헌법/NNG+재판소장/NNG]
편찬위원회장(편찬/NNG+위원회/NNG+장)	[편찬/NNG+위원회장/NNG]
영등포노인종합복지관장	[영등포노인종합복지관장/NNG]
인사부장	[인사부장/NNG]
상황실장	[상황실장/NNG]
비대위원장(비대위/NNG+원+장)	[비대위원장/NNG]

→ 이 단어는 '비대위'에 구성원을 나타내는 접사 '-원', 그리고 우두머리를 나타내는 접시 '-장'이 차례로 결합한 것으로 파악된다. 따라서 '비대위'에 '-원'이 결합하여 '비대위원'이 되고, 여기에 '-장'이 결합하여 '비대위원장'이 된 것으로 보아 전체를 하나의 명사로 분석한다. '영진위원장' 등도 마찬가지이다.

접사가 아니라 '소장, 원장, 지사'와 같은 명사가 결합하여 단체장명이 만들어지는 경우가 있다. 그런 경우에는 아래와 같이 분석한다.

[예시]	헌법재판소소장	[헌법/NNG+재판소/NNG+소장/NNG]
	민족문화연구원원장	[민족/NNG+문화/NNG+연구원/NNG+원장/NNG]
	서울시의원	[서울시/NNP+의원/NNG]
	제주도지사	[제주/NNP+도지사/NNG]

→ '제주도+지사'로 분석할 수도 있으나, 합성어 주의사항 ⑥에 따라 뒤쪽에 더 많은 음절수가 남는 '제주+도지사'를 선택함.

(7) 신문, 잡지, 방송 채널, 웹사이트 등 매체의 이름

매체명이 한 어절에 나타난 경우 통합하여 고유명사로 분석한다. 두 어절 이상으로 분리되어 나온 경우에는 각 어절에 맞는 형태표지를 부여한다. 단, 외국어로 구성된 매체명이 여러 어절로 나타났을 때는 아래 다)의 외국어 처리 지침에 따라 모든 어절을 고유명사로 처리한다.

[예시] 조선일보/NNP, 여성동아/NNP, 폭스뉴스/NNP, 티브이엔/NNP, 보람TV/NNP(개인 채널명)
[예시] 유튜브/NNP, 트위터/NNP, 네이버/NNP, 페이스북/NNP
[예시] 스포츠 서울 [스포츠/NNG, 서울/NNP]
→ 전체 외국어 구성이 아님: 분리된 각 단어에 적합한 형태표지를 부여함.
[예시] 뉴스 위클리 [뉴스/NNP, 위클리/NNP]
스포츠 투데이 [스포츠/NNP, 투데이/NNP]
뉴욕 타임즈 [뉴욕/NNP, 타임즈/NNP]
→ 전체 외국어 구성인 경우: 외국어 지침에 따름

(10) 위에서 명시되지 않은 부류는 모두 고유명사로 인정하지 않는다.

[예시] 임진왜란 (사건명) [임진왜란/NNG]
노벨평화상 [노벨/NNP+평화상/NNG]
네안데르탈인 [네안데르탈인/NNG]
무한도전 (특정 매체의 프로그램 이름) [무한/NNG+도전/NNG]
메이지 (연호) [메이지/NNG]
고양국제꽃박람회 (행사명) [고양/NNP+국제/NNG+꽃/NNG+박람회/NNG]
한국시리즈 (대회명) [한국/NNP+시리즈/NNG]

1.2.2. 언어학적 엄밀성의 추구

① '있다'의 동사 용법과 형용사 용법 구분 기준 제시

<21세기 세종계획>의 문어 형태 분석 말뭉치 구축 지침에는 '있다'의 품사 구분과 관련한 내용이 포함되어 있지 않다. 그리고 이 지침을 바탕으로 실제 구축된 말뭉치에서는 '있다'를 대체로 동사로 처리하였다.

하지만 '있다'는 선어말어미 '-는-' 없이 종결어미 '-다'와 직접 결합하여 쓰이는 경우가

많으므로 형용사로 파악해야 하는 경우가 더 많다. 관형형 어미 '-는'과 결합한다는 점에서는 동사의 모습을 보이나 이때의 '-는'이 '-은'과 대립하며 현재-과거의 구별을 보이지는 않는다는 점을 고려하면 관형형 어미 '-는'과 결합한 '있다'도 동사로 처리할 근거가 약하다.

이를 고려하여 본 사업에서는 '있다'의 대표적인 용법을 형용사 용법으로 파악하는 것이 타당하다고 판단하였으며, 이러한 판단을 바탕으로 '있다'의 동사 용법과 형용사 용법을 구분하는 기준을 제시함으로써 품사 구분의 엄밀성을 추구하고자 하였다.

아래는 '있다'의 용법을 구분하는 방법에 대한 지침의 설명을 보인 것이다. '있다'가 동사로 쓰였다는 적극적인 증거가 있을 때에만 동사로 분석하고, 나머지 경우는 형용사로 분석하도록 하였다. '있다'가 동사로 쓰였음을 보여 주는 적극적인 증거로는 '-는다'와의 결합, '-어라', '-자', '-읍시다' 등 명령형, 청유형 어미와의 결합, '얼마의 시간이 경과하다'의 뜻으로 쓰인 것, '잘'의 수식을 받는 것, '-고 싶다', '-려(고) 하다'와의 결합, '가만히', '마냥'의 수식을 받는 것, 관형형 어미 '-은'과의 결합, 기간을 나타내는 부사어의 수식을 받는 것을 제시하였다.

주의사항

'있다'는 동사 용법과 형용사 용법을 모두 가지고 있다. '있다'는 대개의 경우 형용사로 쓰이는 것으로 보아, '있다'가 동사로 쓰였다는 적극적인 증거가 있을 때에만 동사로 분석하고 나머지 경우는 형용사로 분석한다.

'-고 있다', '-어 있다'형으로 쓰여 앞의 사태가 진행/지속되고 있음을 나타내거나 앞의 사태가 끝나고 그 결과가 유지되고 있음을 나타낸다면 그때의 '있다'는 보조용언(VX)임에 유의한다.

형용사 '있다'의 특징

① '존재하다'의 뜻을 갖는 것은 형용사이다.

[예시] 신이 있다 / 날지 못하는 새도 있다 / 기회가 있다 / 증거가 있다
짜임새 있다 / 쓸모 있다 / 진정성 있다 / 경쟁력 있다 / 가능성 있다 / 필요 있다

② '수 있다', '바 있다', '적이 있다' 구성의 '있다'도 모두 형용사이다.
③ 종결어미 '-다'와 바로 결합하여 '있다.'형으로 쓰이면 형용사이다.

[예시] 이런 경우도 있다. / 그는 서울에 있다.

④ '~에 있어서' 구성의 '있다'도 형용사이다.

[예시] 인간에게 있어서 중요한 것은 사랑이다.

⑤ '누가 어떤 자격으로 있다' 구성의 '있다'도 형용사이다.

[예시] 그는 지금 대기업의 과장으로 있다.
그는 대기 선수로 있다가 출전권을 얻었다.

⑥ 내포문에서 '있다'가 쓰였을 때에는, 종결형으로 바꾸었을 때 '있다.'를 사용하여 표현할 수 있는 경우면 모두 형용사로 판단한다.

[예시] 서울광장에서 있었던 콘서트가 그 예이다.
→ '서울광장에서 콘서트가 있다.'로 바꾸어도 문장이 성립하므로, 형용사로 판단한다.
친구와 둘만 있는 상황이 되면...
→ '(나는 지금) 친구와 둘만 있다.'로 바꾸어도 문장이 성립하므로 형용사로 판단한다.

※ '머물다'의 뜻을 갖는다고 해서 모두 동사로 판단하지 않는다. '머물다'의 의미는 아래와 같이 사전에 동사로도, 형용사로도 기술되어 있다. 따라서 위 ⑥에서 언급했듯이 '있다.'형으로 바꾸어도 문장이 성립되면 '머물다'의 뜻이어도 모두 형용사로 판단하기로 한다. '있다.'형으로 바꿀 수 없거나 '-는다', '-어라', '-자'처럼 동사와 결합하는 어미와 함께 나타났을 때에만 동사로 판단한다.

있다1 [I] 동사
「1」 사람이나 동물이 어느 곳에서 떠나거나 벗어나지 아니하고 머물다.
예) 그는 내일 집에 있는다고 했다.
있다1 [II] 형용사
「2」 사람이나 동물이 어느 곳에 머무르거나 사는 상태이다.
예) 그는 한동안 이 집에 있었다.

⑦ '있다'가 기간을 나타내는 부사어와 함께 쓰일 때에는 종결형 '있다.'를 사용하여 표현하는 것이 어색하다. 이런 경우 동사로 판단한다.

[예시] 그는 노쇠해서 이 자리에 오래 있기 힘들다.
→ '그는 이 자리에 오래 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.
1시간가량 조용히 있다가 갑자기 일어나 총을 꺼내 들었다.
→ '그는 1시간가량 조용히 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.
오늘은 덕수궁 지하도에 더 있다 갈게요.
→ '그는 덕수궁 지하도에 더 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.

동사 '있다'의 특징

① '-는다'(평서), '-어라', '-자', '-읍시다' 등(명령, 청유)이 결합한 것은 동사이다.
② '얼마의 시간이 경과하다'의 뜻일 때에는 동사이다.

[예시] 10분 있다 만나자. / 얼마 안 있어 기다리던 시간이 왔다.

③ 아래 예시와 같이 '잘'과 결합한 '있다'는 동사로 판단한다. '잘 계세요'로 치환이 가능하다는 점에서 동사의 행태를 보이기 때문이다. 단, '내 그물이 잘 있나.'에서처럼 주어가 사람이 아닌 경우에는 동사로 판단하지 않는다.

[예시] (헤어질 때) 잘 있어요. / 아버지는 잘 있느냐. / 잘 있었니.
cf) 잘 계세요. / 아버지는 잘 계시느냐. / 잘 계셨습니까.

④ '-고 싶다', '-려(고) 하다'는 주로 동사와 결합하므로, 이와 결합한 '있다'는 동사로 판단한다.

[예시] 나도 그 자리에 있고 싶다.
나도 여기 있으려고 한다.

⑤ '가만히 있-', '마냥 있-'은, '가만있다'가 동사임을 참고하여, 또 '철수가 가만히 있다.', '철수가 마냥 있다.' 같은 표현이 빈번히 쓰이지 않는 것을 고려하여 동사로 판단한다.
단, 종결어미 '-다'가 바로 결합하여 '가만히 있다.', '마냥 있다.'로 쓰였다면 형용사로 판단한다.
⑥ '있은 지', '있은 후' 등에서 나타나는 '있은'의 '있-'은 동사로 판단한다. '-은'이 결합하여 과거를 나타내는 것이 동사의 특성이기도 하고, '-ㄴ 지', '-ㄴ 후' 등도 주로 동사와 결합하여 쓰이기 때문이다.

[예시] 그 일이 있은 지 수일이 지났다.

② 어미의 문맥상의 기능을 고려하여 형태 표지 부여

<21세기 세종계획>의 형태 분석 말뭉치 구축 지침에서는 종결어미와 연결어미를 후행하는 문장 부호에 따라 구분하도록 하였으나, 그보다는 어미의 문맥상의 기능에 따라 종결어미와 연결어미 표지를 부여하도록 하는 것이 보다 언어학적 엄밀성을 추구하는 방법이 된다.

그러나 동일 형태의 어미가 문장의 말미에 쓰였을 때, 그것이 문장을 끝맺는 종결어미 기능으로 쓰인 것인지 아니면 뒤에 올 말이 생략되었을 뿐 연결어미 기능으로 쓰인 것인지를 명확히 판단하기는 쉽지 않다.

이에 새로운 지침에서는 어미의 종결어미 용법과 연결어미 용법을 기능에 따라 구분하여 형태 표지를 부여하도록 하되, <우리말샘>에 기술된 종결어미, 연결어미 용법을 기준으로 하여 구분하도록 하였다. 가령 어미 '-는지'는 <우리말샘>에서 종결어미와 연결어미 모두로 기술되어 있다.

(2) ㄱ. -는지 「001」「어미」(('있다', '없다', '계시다'의 어간, 동사 어간 또는 어미 '-으시-', '-었-', '-겠-' 뒤에 붙어)) 막연한 의문이 있는 채로 그것을 뒤 절의 사실이나 판단과 관련시키는 데 쓰는 연결 어미.

ㄴ. -는지 「002」「어미」(('있다', '없다', '계시다'의 어간, 동사 어간 또는 어미 '-으시-', '-었-', '-겠-' 뒤에 붙어)) 해할 자리나 간접 인용절에 쓰여, 막연한 의문을 나타내는 종결 어미. 뒤에 보조사 '요'가 오기도 한다.

위와 같이 동일 형태의 어미가 종결어미로도 연결어미로도 등재되어 있는 경우에는 각 용법에 대해 기술된 의미와 예문을 살펴 말뭉치에 등장한 어미의 용법과 유사한 것이 무엇인지를 판단하여 형태 표지를 부여하도록 한 것이다. 이에 따라 아래 (3ㄱ)의 '-는지'는 연결어미로, '모르다'의 목적어 자리에서 간접 의문절을 만들며 쓰인 (3ㄴ)의 '-는지'는 종결어미로 분석하게 된다.

(3) ㄱ. 사람들이 얼마나 떠드는지 공부를 할 수가 없었다. [떠들/VV+는지/EC]
ㄴ. 일을 잘하는지 모르겠다. [잘/MAG+하/XSV+는지/EF]

물론 <우리말샘>의 종결어미, 연결어미 용법 기술에는 보완이 필요한 부분도 포함되어 있으며, 따라서 <우리말샘>의 기술에 기대어 종결어미, 연결어미 형태 표지를 부여하는 방법에는 한계점이 있다. 가령 '-으려고'의 경우 <우리말샘>에 연결어미 용법과 종결어미 용법이 모두 등재되어 있는데, '그 많은 것을 다 먹으려고?'와 같이 의심이나 반문을 나타내는 경우에만 종결어미 용법으로 기술되어 있다. 하지만 아래 예에서 쓰인 '-려고' 역시, 의심이나 반문을 나타내지는 않으나 뒤에 생략된 말이 있다기보다 주어의 의도가 무엇인지만을 밝히며 쓰인 종결어미 용법으로 판단할 수 있다.

(4) 나는 오늘 집에 일찍 가려고.

이 사례가 보여 주듯이 <우리말샘>의 기술만으로 어미의 종결어미 용법과 연결어미 용법을 엄밀하게 구분하기는 어렵지만, 작업의 일관성 확보를 위하여 <우리말샘>의 기술을 연결어미와 종결어미 구분의 기준으로 삼기로 하였다. 다만 (4)와 같은 예를 '-려고'의 종결어미 용법으로 포함할 수 있도록 별도의 기술을 추가하였다.

아울러 기존 지침에서는 '종결어미+요'는 통합해서 종결어미로 분석하고 '비종결어미+요'는 통합하지 않고 각각 분석하도록 하였는데, 종결어미와 연결어미를 구분하는 기준

이 달라지면서 이 지침에도 변경이 필요하게 되었다. 이에 <우리말샘>에 통합형으로 등재된 '-어요, -지요' 등은 통합하여 하나의 표지를 부여하고, 그렇지 않은 경우에는 어미와 보조사 '요'를 분리하여 분석하도록 하였다.

위에서 언급한 것과 관련된 지침의 내용을 보이면 아래와 같다.

나) 종결어미(EF)

용언의 어간이나 선어말어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 한 문장을 끝맺는 역할을 한다. 본 지침에서는 <우리말샘>에 따라 종결어미를 구분한다. 다음은 종결어미의 일부 사례이다.

-게	그만한 돈이 있으면 좋게.	[좋/VA+게/EF+./SF]
-ㄴ가	이것이 무엇인가?	[무엇/NP+이/VCP+ㄴ가/EF+?/SF]
-ㄴ걸	이제 시작인걸.	[시작/NNG+이/VCP+ㄴ걸/EF+./SF]
-ㄴ다	이건 말도 안 된다.	[되/VV+ㄴ다/EF+./SF]
(후략)		

주의사항

① '종결어미+요(보조사)'는 <우리말샘>에 등재되어 있는 '어요, 지요, 래요' 등을 제외하고 모두 종결어미와 보조사로 분리하여 분석한다.

[예시] 말씀대로 했는걸요. [하/VV+았/EP+는걸/EF+요/JX+./SF]

⑤ '-려고'는 <우리말샘>에서 의심과 반문의 용법으로만 종결어미 자격을 갖는 것으로 등재되어 있다. 하지만 아래와 같이 뒤에 생략된 말 없이 주어의 의도만을 밝히며 문말에서 쓰이는 '-려고'는 종결어미 용법으로 볼 수 있으므로 종결어미로 분석한다.

[예시] 나는 오늘 집에 일찍 가려고. [가/VV+려고/EF+./SF]
→ '-려고' 뒤에 '생각하다' 정도의 동사가 생략되어 있다. 이때는 주어의 의도가 무엇인지만을 밝히며 문말에서 쓰인 '-려고'로 볼 수 있으며, 종결어미로 분석한

다.

[예시] 나 요즘 매일 운동해. 살 빼려고. [빼/VV+려고/EC+./SF]

→ '-려고' 뒤에 '생각하다'가 생략된 것이 아니며 '운동하다'와 같이 주어의 의도를 실현하기 위한 행동을 나타내는 말이 생략되어 있다. 이때는 '-려고'가 연결어미로 쓰인 것이다.

다) 연결어미(EC)

용언의 어간이나 선어말어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 문장을 종결시키지 못하고 뒤에 오는 절을 연결시켜 주는 어미를 말한다. 본 지침에서는 <우리말 샘>에 따라 연결어미를 구분하는 것을 원칙으로 한다. 다음은 연결어미의 일부 사례이다.

-거나	누가 오거나 알은 체 할 것 없다.	[오/VV+거나/EC]
-건대	내가 보건대, 네 말이 옳다.	[보/VV+건대/EC]
-고	일을 하고 밥을 먹자.	[하/VV+고/EC]
-곤	숙제한 것도 빌려가곤 한다.	[빌리/VV+어/EC+가/VV+곤/EC]
(후략)		

1.2.3. 과도한 분석 지양

본 지침에서는 과도한 분석으로 품사를 부여하기 어려운 요소가 남는 것을 지양하고, 언중이 인식할 수 있는 단위를 대상으로 형태 표지를 부여하고자 하였다.

① 미등재 합성어의 처리

미등재 합성어의 경우 더 작은 요소로 분리했을 때 어근이 남거나 품사를 부여하기

어려운 요소가 남는다면, 해당 요소를 분리하지 않고 앞말 또는 뒷말과 결합하여 형태 표지를 부여하도록 하였다. 가령 '당정청'이라는 합성어가 말뭉치에 자주 등장하는데, 현재 <우리말샘>에 '당정청'은 등재되어 있지 않고 '당정'만이 등재되어 있다. 이런 경우 '당정+청'으로의 분리 분석을 고려할 수 있는데, 청와대를 뜻하는 '청'이 <우리말샘> 미등재어로서 문장 속에서 홀로 쓰이는 일이 드물어 어근의 일종으로 파악된다. 이런 경우 '당정/NNG+청/NNG'으로 분석하여 언중이 인식하기 어려운 단위를 따로 분리하는 것보다 '당정청/NNG'으로 분석하는 것이 직관에도 부합하고 실용적으로도 장점이 있다. 이러한 점을 고려하여 미등재 합성어 처리 지침을 아래와 같이 제시하였다.

주의사항

② 어절에 나타난 표기가 규범에 맞지 않아 사전에서 검색되지 않으나 규범에 맞게 표기된 단어는 사전 등재어일 때, 규범에 맞게 표기된 단어에 준하여 한 단어로 분석한다.

[예시] 먼저번 (사전: 먼젓번) [먼저번/NNG]
조랭이떡 (사전: 조롱이-떡) [조랭이떡/NNG]

③ 단어 자체가 사전의 표제어로 등록되어 있지는 않으나 사전에 구로 등재되어 있는 말(A^B)의 일부에 해당하는 단어일 때에도 한 단위로 분석한다.

[예시] 사대강을 (사전: 사대강^수계법) [사대강/NNG+을/JKO]

⑤ 아래와 같은 혼성어는 분리하여 분석하기 어려우므로 한 단어로 보고, 의미에 따라 일반명사 또는 고유명사로 분석한다.

[예시] 아베노믹스 (아베+이코노믹스) [아베노믹스/NNG]
→ 이는 아베의 경제정책을 일컫는 말로, 의미상 본 지침의 고유명사 부류에 들지 않는다. 이에 따라 전체 단어를 일반명사로 처리한다.
[예시] 홍드로 (홍수아+페드로) [홍드로/NNP]
→ 이는 특정 인물을 가리키는 말로 사용되고 있으므로 의미상 고유명사 부류에 든다. 이에 따라 전체 단어를 고유명사로 처리한다.

⑥ 합성어로 등록되어 있지 않은 표제어는 분리해서 분석하되, 사전 표제어로 등록되어 있는 최대한 많은 음절수의 단어를 생성하도록 나눈다. 즉 다음 예와 같은 경우 3음절 어휘가 생성되는 첫 번째 분석을 취한다.

[예시]	영상학과	[영상학/NNG+과/NNG] (3음절+1음절)
	영상학과	[영상/NNG+학과/NNG] (2음절+2음절)

⑦ 3음절 어휘와 같이 어느 쪽으로 나뉘어도 음절수가 같고, 양쪽 분석이 모두 사전 표제어라면 뒤쪽을 먼저 분석한다.

[예시]	차창밖	[차/NNG+창밖/NNG]
	이등품	[이/NR+등품/NNG]

⑧ 합성어로 등록되어 있지 않은 표제어를 더 작은 요소로 분리했을 때 어근이 남거나 품사를 부여하기 어려운 요소가 남는다면, 해당 요소를 분리하지 않고 앞말 또는 뒷말과 결합하여 형태 표지를 부여한다.

[예시]	당정청	[당정청/NNG]

사전에는 '당정청'이 올라 있지 않고 '당정'만이 올라 있다. 청와대를 뜻하는 '청'은 미등재어이고 홀로 쓰이는 일이 드물어 어근으로 판단할 수 있다. 이런 경우 '청'을 앞말인 '당정'과 결합하여 처리한다.

[예시]	오인승	[오인승/NNG]

'오인승'은 수사 '오'와 명사 '인', 그리고 사전 미등재어인 '승'으로 구성되어 있다. '승'은 미등재어이지만 홀로 쓰이지 않으므로 어근으로 판단할 수 있고, 이런 경우 '승'을 앞말인 '인'과 결합하여 처리한다. 그런데 그렇게 해서 도출된 '인승' 역시 미등재어이고, '인승'의 '인'이 일반명사임을 고려하면 '인승'도 일반명사가 되어야 할 것이나 이 말이 홀로 쓰이지 않기 때문에 일반명사로 품사를 부여하기가 어렵다. 따라서 '인승'도 분리하지 않고 앞말과 결합하여 형태 표지를 부여한다.

② 기타 문자가 포함된 어절의 처리

기타 문자가 포함된 어절의 경우 한글과 기타 문자를 분리했을 때 다른 품사로 분석될 수 없는 단위가 남는다면, 분리하지 않고 통합하여 형태 표지를 부여하도록 하였다. 가령 '1루수'의 경우 숫자 '1'을 제외하면 '루수'라는 단위가 남는데, 두음법칙에 따라 '누수(壘手)'는 명사로 볼 수 있고 <우리말샘>에도 등재되어 있지만 '루수'는 그 자체로 명사로 보기가 어렵고 <우리말샘>에도 등재되어 있지 않다. 이 경우 기타 문자가 포함되어 있다는 이유로 '1'과 '루수'를 분리하면, 다른 품사로 분석될 수 없고 언중이 인식하기 어려운 단위가 과도하게 분리되는 결과를 낳게 된다. 이러한 점을 고려하여 기타 문자를 분리했을 때 다른 품사로 분석될 수 없는 단위가 남는다면 기타 문자와 한글을 분리하지 않고 통합하여 형태 표지를 부여하도록 한 것이다. 이에 따르면 '1루수'는 '1루수/NNG'로 처리된다. 이와 관련된 지침을 자세히 보이면 아래와 같다.

가) 일반명사(NNG)

(1) 일반명사로 분석할 수 있는 단어

(라) 기타 다른 품사로 분석될 수 없는 단위

표기상 한글과 기타 문자(부호나 숫자, 외국 문자)가 섞여 있고, 한글과 기타 문자를 분리했을 때 '절', '루수', '관왕', '유'같이 다른 품사로 분석될 수 없는 단위가 도출되는 경우에는 분리하지 않고 통합하여 분석한다.

[예시]	3.1절(국경일)	[3.1절/NNG]
	1루수	[1루수/NNG]
	5관왕	[5관왕/NNG]
	벙커C유	[벙커C유/NNG]

기타 문자가 포함된 단위에서 접사를 분리할 수 있을 것으로 생각되더라도, 접사 분리 시 단어의 의미 구조와 맞지 않게 된다면 접사 분리를 하지 않는다.

[예시] 제3자 [제3자/NNG]

→ '제-'는 본 지침의 분석 대상 접사이고, '-자'는 분석 대상이 아닌 접사이다. 이때 '제-'를 '3자'로부터 분리하여 [제/XPN+3자/NNG]로 분석하면, 이 단어의 의미 구조가 '제3의 사람', 즉 '제3-자'인 것과 맞지 않게 된다. 따라서 접사 '제-'를 분리하지 않고 전체 단어를 일반명사로 처리한다.

[예시] 제3국 [제3국/NNG]

→ 위와 마찬가지로 '제-'는 본 지침의 분석 대상 접시아고 '-국'은 분석 대상이 아닌 접사이다. 역시 '제-'를 '3국'으로부터 분리하여 [제/XPN+3국/NNG]로 분석하면, 이 단어의 의미 구조가 '제3의 나라', 즉 '제3-국'인 것과 맞지 않게 된다. 따라서 접사 '제-'를 분리하지 않고 전체 단어를 일반명사로 처리한다.

기타 문자를 분리하고 남는 단위가 '의존명사+비분석 접사'인 경우에는, '의존명사+비분석 접사'를 합하여 의존명사로 처리한다.

[예시] 16개교 [16/SN+개교/NNB]

→ '개'는 의존명사이고 '교'는 사전에 등재되어 있지 않으나 접사에 준하는 요소로 파악된다. 또한 이 '-교'는 본 지침에서 분석하지 않는 접사이다. 이때 비분석 접사인 '-교'를 의존명사 '개'에 합하여 의존명사 '개교'를 설정하여 분석한다.

[예시] 16강전 [16/SN+강전/NNB]

→ 위의 경우와 마찬가지로 의존명사 '강'에 비분석 접사 '-전'이 결합한 '강전'을 의존명사로 설정하여 분석한다. <우리말샘>에 '16강전'이 '십육-강전'이 아닌 '십육강-전'으로 올라 있어 본 지침의 처리와 달라지기는 하나, '개교'와의 구조적 유사성, '강전'이 다양한 수사와 어울려 쓰이며 의존명사와 같은 행태를 보임에 주목하는 것이다.

[참고] 십육강전 [십육강전/NNG]

→ '십육강전'이 기타 문자 없이 한글만으로 쓰인 경우에는 사전의 처리를 따라 '십육강전/NNG'으로 분석한다. 기타 문자를 이용해서 표기했는지 한글만으로 표기했는지에 따라 달리 처리하게 되지만, 기타 문자는 분리하는 것이 원칙이라는 점을 고려한 것이다.

단, 숫자나 외국어로만 표기된 경우에는 모두 각각을 분석한다.

[예시] 6.25 [6/SN+./SP+25/SN]

③ 부호가 개입한 어절의 처리

부호가 개입한 말의 경우, 부호를 뺀 말이 사전에 한 단어로 등재되어 있다면 전체를 통합하여 형태 표지를 부여하도록 하였다. 가령 '농・수산물'처럼 한 어절 안에 가운뎃점이 포함되어 있을 때, 가운뎃점이 포함되어 있다는 이유로 각 요소를 분리하여 분석하면 다른 품사로 분석될 수 없고 언중이 인식하기 어려운 단위인 '농'이 과도하게 분리되는 결과를 낳게 된다. 이러한 점을 고려하여 부호를 뺀 말이 사전에 한 단어로 등재되어 있다면 전체를 통합하여 형태 표지를 부여하도록 한 것이다. 이에 따르면 '농・수산물'은 '농・수산물/NNG'로 처리된다. 이와 관련된 지침을 자세히 보이면 아래와 같다.

바) 기호가 어절 중간에 개입한 경우

기호가 어절 중간에 개입한 경우, 기호를 뺀 말이 사전에 한 단어로 등재되어 있다면 기호가 있다 하더라도 전체를 통합하여 표지를 부여한다.

[예시] 농·수산물 (사전: 농수산-물)	[농·수산물/NNG]
초·중·고 (사전: 초중고)	[초·중·고/NNG]
의~리	[의~리/NNG]
사이~소 (사전: 어미 '-이소')	[사/VV+이~소/EF]

기호를 뺀 말이 사전에 한 단어로 등재되어 있지 않은 경우에도, 분리하여 분석할 경우 어근이 남는다면 전체를 통합하여 표지를 부여한다.

[예시] 당·정·청 (사전: 당정)	[당·정·청/NNG]

사전에 '당정'만이 등재되어 있어 이 어절을 '당·정/NNG'과 '·/SP', '청'으로 분리할 경우, 사전 미등재어이면서 홀로 쓰이지 않는 '청'이 남는다. 이 경우 '청'을 앞말에 통합하여 '당·정·청/NNG'로 표지를 부여한다.

단, 숫자나 외국어로만 표기된 경우에 기호가 포함되어 있으면 모두 각각 분석한다.

[예시] 6.25	[6/SN+./SP+25/SN]

④ 외국어의 처리

국제적 교류가 늘어나면서 우리가 사용하는 언어 속에도 다량의 외국어 요소가 들어오게 되었다. 우리말 속에 들어온 외국어 요소 중에는 우리말 문장을 구성할 때 단독으로 쓰일 수 있어 단어의 자격을 갖는 것이 있는가 하면, 단독으로는 잘 쓰이지 않고 주로 다른 말과 결합해서 쓰여 단어 자격을 갖지 못하고 의미 요소로서의 역할만을 하는 것이 있다. 가령 '그룹(group)'이라는 외국어 요소는 '그룹을 지어 모여라'에서처럼 단어 자격을 가지고 우리말 문장 속에서 쓰이며 <우리말샘>에도 명사로 등재되어 있다. 이에 비해 '걸(girl)'이라는 외국어 요소는 '저기 한 *걸이 지나간다'처럼 쓰이지 않아 우리말에서 단어 자격을 갖는다고 보기 어려우며 <우리말샘>에도 등재되어 있지 않다.

그러면 이 두 단어가 만나 이루어진 말인 '걸그룹'은 어떻게 분석할 수 있을지에 대해 생각해 보자. 이 말은 한국인의 일상에서 흔히 접할 수 있는 말에 해당하며, <우리말샘>에는 '걸 그룹'으로, 즉 띄어 쓴 표제어로 올라 있다. 이처럼 사전에 '구'로 등재되어 있는 말은 구를 이루는 요소들을 분리하여 분석하도록 해 왔으나, '걸그룹'을 '걸+그룹'으로 분리하여 분석할 경우 한국어에서 단어 자격을 갖는다고 보기 어려운 '걸'을 명사로 취급하게 된다는 점에서 문제가 생긴다.

이에 한국어에서 단어 자격을 가질 수 있는 언어 단위를 분석 대상으로 삼기 위하여 외국어 처리 지침을 새로이 마련하였다. 아래에서 외국어 처리의 기본 원칙을 소개한다.

첫째, 어절 속에 포함된 외국어 요소가 한 단어라면, <우리말샘> 등재 여부와 무관하게 그 단어를 의미에 따라 고유명사 또는 일반명사로 분석한다.

(5) 마이너 리그 [마이너/NNG, 리그/NNG]

둘째, 어절 속에 포함된 외국어 요소가 둘 이상의 단어일 때에는 그 외국어 요소를 단어 단위로 분리한 후, 분리된 각 단어가 고유명사 또는 일반명사로 따로따로 처리될 수 있으면 각 단어를 분리하여 형태 표지를 부여한다. (어떤 단어가 <우리말샘>에 단독으로 일반명사로 등재되어 있으면 그 단어는 일반명사 처리가 가능하다고 본다.)

(6) 배팅글러브 ('배팅' 등재, '글러브' 등재) [배팅/NNG+글러브/NNG]

셋째, 분리된 각 단어 중 어느 하나라도 고유명사 또는 일반명사로 처리될 수 없다면 전체를 묶어서 형태 표지를 부여한다.

(7) 걸그룹 ('걸' 미등재, '그룹' 등재) [걸그룹/NNG]

한편, 외국어의 한 문장이 한글로 전사되어 나타난 경우에는 각 어절을 분석불능범주(NA)로 처리하도록 하였다. 해당 문장 속의 언어 요소들이 국어의 요소로서 사용된 것이라 볼 수 없기 때문이다.

(8) 하우 두 유 두? [하우/NA, 두/NA, 유/NA, 두/NA+?/SF]

이에 더해 두 어절 이상의 외국어로 이루어진 고유명사의 경우 각 어절 모두를 고유명사로 처리하도록 하였다. 본 사업의 형태 분석은 '어절' 단위의 분석을 원칙으로 하므로 두 어절 이상으로 구성된 고유명사(예: 바람과 함께 사라지다)는 고유명사로 처리하지 않고 각 어절을 품사에 맞게 분석하는 방침을 따르고 있으나, '블루 이즈 더 워미스트 컬러'(영화 제목)처럼 두 어절 이상의 외국어로 이루어진 고유명사에 대해 각 어절을

일반명사로 처리하면, 고유명사 정보가 사라질 뿐 아니라 국어에서 유의미하게 사용되는 언어 단위가 아닌 것이 일반명사로 처리된다는 점에서도 가치가 낮은 정보를 산출하게 된다. 이를 고려하여 외국어로 구성된 두 어절 이상의 고유명사에는 예외 조항을 두어 각 어절 모두를 고유명사로 처리하기로 한 것이다.

(9) 블루 이즈 더 워미스트 컬러(영화 제목)

[블루/NNP, 이즈/NNP, 더/NNP, 워미스트/NNP, 컬러/NNP]

이러한 지침을 토대로 외국어 요소 중 한국어에서 단어 자격을 가지는 것으로 인식되는 단위를 분석 대상으로 삼고자 하였다. 위의 내용을 담은 외국어 처리 지침은 아래와 같다.

다) 외국어의 처리

외국어는 아래의 방식에 따라 처리한다.

(1) <우리말샘>에 한 단어로 등재된 외국어

전체를 묶어 의미에 따라 NNG 또는 NNP로 처리한다. 그 외의 경우(<우리말샘>에 등재되지 않았거나 구로 등재된 것, <우리말샘>에 한 단어로 등재되었지만 원문에서 여러 어절로 분리되어 나타난 것)에 대한 처리는 아래의 (2)에 따른다.

[예시] 가든파티 (사전: 가든-파티)	[가든파티/NNG]
[예시] 마추픽추 (사전: 마추픽추)	[마추픽추/NNP]

(2) <우리말샘>에 등재되지 않았거나 구로 등재된 외국어

(가) 한 어절로 나타났든 두 어절 이상으로 나타났든 **전체 외국어 표현이 본 지침의 고유명사 부류에 든다면,** 그 고유명사를 이루는 **각 어절 모두를 (어절 내부 분석 없이) NNP로** 처리한다.

[예시] 레드제플린 (록 그룹 이름)	[레드제플린/NNP]
레드 제플린	[레드/NNP, 제플린/NNP]
[예시] 카미노데산티아고 (지명, 순례길 이름)	[카미노데산티아고/NNP]
카미노 데 산티아고	[카미노/NNP, 데/NNP, 산티아고/NNP]
[예시] 로버트다우니주니어 (인명)	[로버트다우니주니어/NNP]
로버트 다우니 주니어	[로버트/NNP, 다우니/NNP, 주니어/NNP]
페르디낭 드 소쉬르	[페르디낭/NNP, 드/NNP, 소쉬르/NNP]
cf) 루이9세	[루이/NNP+9/SN+세/NNB]
루이 9세	[루이/NNP, 9/SN+세/NNB]
[예시] 웻 앤 와일드 워터월드 (시설물명)	[웻/NNP, 앤/NNP, 와일드/NNP, 워터월드/NNP]
[예시] 블루 이즈 더 워미스트 컬러 (영화제목)	[블루/NNP, 이즈/NNP, 더/NNP, 워미스트/NNP, 컬러/NNP]

(나) 위의 경우가 아니라면, 외국어를 포함하고 있는 각각의 어절에 대하여 다음과 같은 절차를 적용하여 처리한다.

(나-1) **각각의 어절 속에 포함된 외국어 요소가 한 단어라면,** <우리말샘> 등재 여부와 무관하게 그 단어를 의미에 따라 **고유명사 또는 일반명사로** 분석한다.

[예시] 마이너하다	[마이너/NNG+하/XSA+다/EF]
→ 미등재어인 '마이너'는 '마이너 감성', '메이저와 마이너의 경계' 등에서 볼 수 있듯이 다른 말과도 결합하여 쓰이므로 어근보다는 명사의 성격을 띠는 단어로 볼 수 있다. 외국어의 이런 특성을 고려하여, 한 단어에 해당하는 미등재 외국어를 어근(XR)이 아닌 체언류로 처리한다.	
[예시] 마이너 리그	[마이너/NNG, 리그/NNG]
[예시] 라 리가 (축구 리그)	[라/NNG, 리가/NNG]

(나-2) **한 어절 속에 포함된 외국어 요소가 둘 이상의 단어일 때에는** 그 외국어 요소를 **단어 단위로 분리한다.** 단어 단위를 판단할 때에는 해당 외국어에서의 표기법(띄어쓰기 여부)을 참고한다. 그 후 아래의 지침을 따른다.

[예시] 배팅글러브 [배팅, 글러브]
[예시] 아시안게임 [아시안, 게임]
[예시] 리우올림픽 [리우, 올림픽]
[예시] 보이그룹 [보이, 그룹]
[예시] 라리가 (축구 리그) [라, 리가]

(나-3) 분리된 각 단어가 본 지침의 고유명사 부류에 들거나 <우리말샘>에 단독 일반명사로 등재되어 있어서 **각각의 단어를 따로 처리할 수 있는 상황이라면, 각 단어를 분리하여 형태표지를 부여**한다.

[예시] 배팅글러브 ('배팅' 등재, '글러브' 등재) [배팅/NNG+글러브/NNG]
[예시] 아시안게임 ('아시안' 고유명사, '게임' 등재) [아시안/NNP+게임/NNG]
→ '아시안', '아메리칸', '브리티시' 등 고유명사의 형용사형을 모두 고유명사로 처리한다.
[예시] 리우올림픽 ('리우' 고유명사, '올림픽' 등재) [리우/NNP+올림픽/NNG]

(나-4) 분리된 각 단어 중 어느 하나라도 위의 방식에 따라 고유명사로 또는 일반명사로 **처리할 수 없다면,** 각 단어를 분리하지 않고 **전체를 묶어서 의미에 따라 고유명사 또는 일반명사로** 분석한다.

[예시] 보이, 그룹: '그룹'은 팀의 의미로 단독으로 등재되어 있으나, '보이'는 '소년'의 의미로서는 단독으로 명사로 등재되지 않음. '보이'의 처리가 어려우므로 전체를 묶어 일반명사로 분석함.
[보이그룹/NNG]
[예시] 라, 리가: '라'와 '리가' 모두 고유명사 또는 일반명사로 처리하기 어려움. 대회명은 본 지침에서 고유명사에 들지 않으므로 전체를 묶어 일반명사로 분석함.
[라리가/NNG]

(나-5) 위에 제시한 절차를 외국어를 포함하고 있는 모든 어절에 각각 적용한다. 예시는 다음과 같다.

[예시] 피지컬 트레이닝	[피지컬/NNG, 트레이닝/NNG]
[예시] 워터해저드	[워터해저드/NNG]
[예시] 골든 커리어 그랜드슬램 (기록명)	[골든/NNG, 커리어/NNG, 그랜드슬램/NNG]
[예시] 글로벌 북카페 (신문 코너명)	[글로벌/NNG, 북카페/NNG]
글로벌 북 카페	[글로벌/NNG, 북/NNG, 카페/NNG]

(다) 아래와 같이 **외국어의 '한 문장'이 한글로 전사되어 나타난 경우, 각 어절을 내부 분석 없이,** 그리고 각 단어의 <우리말샘> 등재 여부와 무관하게 **NA로** 처리한다.

[예시] 렛츠고	[렛츠고/NA]
[예시] 익스큐즈 미	[익스큐즈/NA, 미/NA]
[예시] 아이 러브 유	[아이/NA, 러브/NA, 유/NA]
[예시] 굿!	[굿/NA+!/SF]
[예시] 곤니치와	[곤니치와/NA]
[예시] 니하오	[니하오/NA]
[예시] 해피버스데이 투 유	[해피버스데이/NA, 투/NA, 유/NA]

1.2.4. 기타

여기에서는 위에서 언급한 사항 외에 <21세기 세종계획>의 지침과 달라진 주요 내용을 보이기로 한다.

① 문어와 구어 분석 지침의 통합

본 사업에서 마련한 지침은 문어뿐 아니라 구어 분석 시에도 적용할 수 있도록 구성되었다. 즉 문어와 구어는 원칙적으로 동일한 형태 표지 목록을 바탕으로 동일한 방식으로 분석된다.

다만 구어는 문어와 달리 다양한 준말과 형태 변이 현상을 보여 주므로 구어에서 나

타나는 준말과 형태 변이 현상의 처리 방법을 따로 명시할 필요가 있다. 또한 구어 전사 시에 이용된 특별한 마크업과 표지가 있기에, 그것의 처리 방법도 별도로 제시할 필요가 있다.

이에 지침의 말미에 구어 분석 시의 유의점에 대한 지침을 붙였는데, 본 사업에서는 아래 예시와 같이 구어에서 나타나는 준말과 형태 변이 현상을 되도록 분석에 반영하는 방향으로 지침의 내용을 마련하였다.

(10) ㄱ. 건(<그건) 어렵지 않아요 [거/NP+ㄴ/JX]
ㄴ. 늦을까 봐 날라서 왔어. [날르/VV+아서/EC]
ㄷ. 좋으까? [좋/VA+으까/EF+?/SF]
ㄹ. 갈 것 같애 [같/VA+애/EF]
ㅁ. 이케 [잏/VA+게/EC]
ㅂ. 그치 않습니까? [귷/VA+지/EC]
ㅅ. 내비둬 [내비두/VV+어/EF]
ㅇ. 언놈이(<어느 놈이) 그래? [언놈/NP+이/JKS]
ㅈ. 짱난다 [짱나/VV+ㄴ다/EF]

이 중 (10ㅁ, ㅂ)은 용언 어간과 어미가 축약된 형식을 분석할 때 어간을 '잏', '귷'과 같이 복원하도록 하였음을 보여 준다. 어간이 낯선 형식으로 분석되었다는 단점이 있지만, 문장에서 용언이 중요한 역할을 하는 만큼 향후 구문 분석 말뭉치 구축 등을 위하여 용언 어간을 어미와 분리하여 드러낼 필요가 있었고, 그러면서도 구어의 준말 현상을 드러낼 수 있다는 점에 주목하여 이와 같은 분석 방식을 채택하였다.

② 관형사의 세분

<21세기 세종계획>의 지침에서는 관형사의 하위 부류를 구분하지 않았으나, 본 지침에서는 관형사를 의미에 따라 지시관형사, 수관형사, 성상관형사로 세분하여 형태 표지를 부여하였다.

의미상 직시의 성격이 있는 관형사를 지시관형사로 포함하고, 수량이나 차례를 나타내는 관형사를 수관형사로 포함하며, 그 외의 관형사는 성상관형사로 포함하였다.

(11) 지시관형사: 이/MMD, 그/MMD, 저/MMD, 어느/MMD, 아무/MMD, 현/MMD, 오른/MMD
수관형사: 한/MMN, 두/MMN, 여러/MMN, 모든/MMN, 양/MMN
성상관형사: 새/MMA, 구/MMA, 약/MMA

③ 분석 대상 접사의 목록

본 사업에서는 생산성이 높은 접사를 중심으로 하여 일부 접두사, 접미사만을 분리 분석 대상으로 삼았다. 그 목록은 <21세기 세종계획>에서 제시한 목록과 거의 동일하며 접두사 목록에서는 변화가 없다. 다만 접미사 목록에는 일부 변화가 있다. 우선 분석 대상 명사파생접미사 목록에 주로 구 단위에 결합하는 접미사를 중심으로 다음 아홉 개의 접미사를 추가하였다.

(12) 추가된 명사파생접미사

가량	1시간가량, 다섯명가량
간(間)	한 달간
권(券)	만 원권

발(發)　　서울발
부(附)　　12일부
분지(分之)　삼분지 일
어치　　만 원어치
정(整)　　일만 원정
하(下)　　지배하

또한 분석 대상 동사파생접미사 목록에 피동의 뜻을 더하는 '-받-'을 추가하였다.

(13) 추가된 동사파생접미사
　　받　　집세 인상을 강요받았다.

④ 유물명, 식물명, 동물명

'청자상감국화무늬긴목병'과 같은 유물명, '북부점박이올빼미'와 같은 동물명 등은 세분하지 않고 전체를 묶어 일반명사로 분석하도록 하였다. 이들이 하나의 대상을 지시한다는 점을 고려한 것이다.

⑤ 의존명사 '것'과 '거'

기존 지침에서는 다른 형태와의 결합에서 '거'의 형태가 유지되지 않는다면 '것'으로 복원하여 분석하도록 하였고, 이에 따라 '걸, 건, 게'가 아래와 같이 분석되었다.

(14) ㄱ. 공부할 걸 가져왔니?　[것/NNB+ㄹ/JKO]
　　ㄴ. 연습할 건 있니?　　[것/NNB+ㄴ/JX]

ㄷ. 먹을 게 모자라다.　[것/NNB+이/JKS]

그러나 '거'의 형태를 그대로 인정하는 것이 직관에 부합하므로, 아래와 같이 '거'의 형태를 반영하여 분석하도록 하였다.

(15) ㄱ. 공부할 걸 가져왔니?　[거/NNB+ㄹ/JKO]
ㄴ. 연습할 건 있니?　[거/NNB+ㄴ/JX]
ㄷ. 먹을 게 모자라다.　[거/NNB+이/JKS]

⑥ 'ᆭ'

'저 오늘 일찍 일어났ᆭ아요'에서처럼 "앞의 사실을 청자가 이미 알고 있음"을 나타내는 'ᆭ은, 선어말어미처럼 취급하여 어말어미와 결합하여 표지를 부여하도록 하였다. 이때의 'ᆭ을 '지+않'으로 분석하기에는 분포와 기능이 뚜렷하게 변하였다는 점을 고려한 것이다.

(16) 저 오늘 일찍 일어났ᆭ아요.　[일어나/VV+았/EP+ᆭ아요/EF+./SF]

⑦ 마침표

마침표(.)는 문장 종결부에서 쓰이는 경우가 많지만 문장 종결부가 아닌 위치에서도 자주 쓰이며 여러 가지 기능을 한다. 이에 문장 끝에서 쓰인 마침표는 SF로, 소수점으로 쓰이거나 날짜를 나타내는 숫자 사이에서 쓰인 마침표, 홈페이지 주소 속에서 쓰인 마침표 등 문장 종결의 의미가 없는 것은 SP로 분석하도록 하였다. 또한 말줄임표 대신 마침표가 쓰인 경우에는 마침표를 모두 묶어 SE로 분석하도록 하였다.

2. 형태 분석 말뭉치 구축 지침

이 장에서는 본 사업에서 형태 분석 말뭉치를 구축하기 위하여 적용한 지침의 전모를 보이고자 한다. 지침의 구성은 다음과 같다.

1. 기본 원칙
   가. 분석대상
   나. 분석원리
   다. 분석원칙
2. 어절 분석 표지
3. 표지별 분류 기준 및 세부 지침
   가. 체언
   나. 용언
   다. 수식언
   라. 독립언
   마. 관계언
   바. 의존형태
   사. 기타
   아. 구어

1 기본 원칙

가. 분석대상

형태 분석은 하나의 어절을 분석 대상으로 한다.

나. 분석원리

본 분석은 ‘형태소’ 차원이 아닌 ‘형태’ 차원의 분석이므로 이형태를 최대한 반영한다.

다. 분석원칙

형태분석은 분석 대상인 원시 말뭉치를 가급적 훼손하지 않는다. 본 분석은 국립국어원의 <우리말샘>의 표제어를 기준으로 한다.

2. 어절 분석 표지

가. 이 어절 분석표지는(이하 세종 형태 표지) 21세기 세종계획 국어기초자료 구축 분과에서 '형태 분석 말뭉치(morpheme tagged corpus)'를 구축하기 위해 마련된 것을 토대로 작성된 것이다.

나. 이 분석 표지는 큰 틀은 21세기 세종계획의 어절 분석 표지를 따르고, 품사 태그의 경우는 TTA의 분석 표지를 참고하였다. 그리고 세종 말뭉치의 문어, 구어 분석 표지를 통합한 것이다.

다. 이 형태 표지는 단계적인 분석을 할 수 있도록 고려하였다.

대분류	소분류	세분류	태그
체언	명사	일반명사	NNG
		고유명사	NNP
		의존명사	NNB
	대명사	대명사	NP
	수사	수사	NR
용언	동사	동사	VV
	형용사	형용사	VA
	보조용언	보조용언	VX
	지정사	긍정지정사	VCP
		부정지정사	VCN
수식언	관형사	성상 관형사	MMA
		지시 관형사	MMD
		수 관형사	MMN
	부사	일반부사	MAG
		접속부사	MAJ
독립언	감탄사	감탄사	IC

관계언	격조사	주격조사	JKS
		보격조사	JKC
		관형격조사	JKG
		목적격조사	JKO
		부사격조사	JKB
		호격조사	JKV
		인용격조사	JKQ
	보조사	보조사	JX
	접속조사	접속조사	JC
의존형태	어미	선어말어미	EP
		종결어미	EF
		연결어미	EC
		명사형전성어미	ETN
		관형형전성어미	ETM
	접두사	체언접두사	XPN
	접미사	명사파생접미사	XSN
		동사파생접미사	XSV
		형용사파생접미사	XSA
	어근	어근	XR
기호	일반기호	마침표, 물음표, 느낌표	SF
		쉼표, 가운뎃점, 콜론, 빗금	SP
		따옴표, 괄호표, 줄표	SS
		줄임표	SE
		붙임표(물결)	SO
		기타 기호	SW
	외국어	외국어	SL
	한자	한자	SH
	숫자	숫자	SN
	분석불능범주	분석불능범주	NA
		명사추정범주	NF
		용언추정범주	NV

3 표지별 분류 기준 및 세부 지침

가 체언

체언은 명사, 대명사, 수사를 포괄하는 대범주로서, 조사와 결합하거나 그 자체로 다른 체언이나 용언과 어울려 하나의 문장성분이 될 수 있다.

1) 명사(NN)

명사는 사물의 이름을 나타내는 품사이다. 본 표지에서는 명사를 일반명사, 고유명사, 의존명사로 세분한다.

가) 일반명사(NNG)

사물의 이름을 나타내는 단어로서 <우리말샘>에 명사로 등재된 표제어(고유명사와 의존명사를 제외한 모든 명사)와 독립된 음절(한자어), 약어, 고사성어 등 사전 표제어는 아니나 다른 품사로 분석될 수 없는 단위들을 포함한다.

(1) 일반명사로 분석할 수 있는 단어

(가) <우리말샘>의 명사 표제어

[예시] 국어/NNG, 연구/NNG

(나) 1음절 한자어가 독립된 단위로 사용되는 경우

[예시] 서울초등학교 졸	[졸/NNG]

(다) 한자성어

[예시] 백척간두(百尺竿頭)	[백척간두/NNG+(/SS+百尺竿頭/SH+)/SS]

(라) 기타 다른 품사로 분석될 수 없는 단위

표기상 한글과 기타 문자(부호나 숫자, 외국 문자)가 섞여 있고, 한글과 기타 문자를 분리했을 때 '절', '루수', '관왕', '유'같이 다른 품사로 분석될 수 없는 단위가 도출되는 경우에는 분리하지 않고 통합하여 분석한다.

[예시]	3.1절(국경일)	[3.1절/NNG]
	1루수	[1루수/NNG]
	5관왕	[5관왕/NNG]
	벙커C유	[벙커C유/NNG]

기타 문자가 포함된 단위에서 접사를 분리할 수 있을 것으로 생각되더라도, 접사 분리 시 단어의 의미 구조와 맞지 않게 된다면 접사 분리를 하지 않는다.

[예시] 제3자 [제3자/NNG]

→ '제-'는 본 지침의 분석 대상 접사이고, '-자'는 분석 대상이 아닌 접사이다. 이때 '제-'를 '3자'로부터 분리하여 [제/XPN+3자/NNG]로 분석하면, 이 단어의 의미 구조가 '제3의 사람', 즉 '제3-자'인 것과 맞지 않게 된다. 따라서 접사 '제-'를 분리하지 않고 전체 단어를 일반명사로 처리한다.

[예시] 제3국 [제3국/NNG]

→ 위와 마찬가지로 '제-'는 본 지침의 분석 대상 접사이고 '-국'은 분석 대상이 아닌 접사이다. 역시 '제-'를 '3국'으로부터 분리하여 [제/XPN+3국/NNG]로 분석하면, 이 단어의 의미 구조가 '제3의 나라', 즉 '제3-국'인 것과 맞지 않게 된다. 따라서 접사 '제-'를 분리하지 않고 전체 단어를 일반명사로 처리한다.

기타 문자를 분리하고 남는 단위가 '의존명사+비분석 접사'인 경우에는, '의존명사+비분석 접사'를 합하여 의존명사로 처리한다.

[예시] 16개교 [16/SN+개교/NNB]

→ '개'는 의존명사이고 '교'는 사전에 등재되어 있지 않으나 접사에 준하는 요소로 파악된다. 또한 이 '-교'는 본 지침에서 분석하지 않는 접사이다. 이때 비분석 접사인 '-교'를 의존명사 '개'에 합하여 의존명사 '개교'를 설정하여 분석한다.

[예시] 16강전 [16/SN+강전/NNB]

→ 위의 경우와 마찬가지로 의존명사 '강'에 비분석 접사 '-전'이 결합한 '강전'을 의존명사로 설정하여 분석한다. <우리말샘>에 '16강전'이 '십육-강전'이 아닌 '십육강-전'으로 올

라 있어 본 지침의 처리와 달라지기는 하나, '개교'와의 구조적 유사성, '강전'이 다양한 수사와 어울려 쓰이며 의존명사와 같은 행태를 보임에 주목하는 것이다.

[참고] 십육강전 [십육강전/NNG]

→ '십육강전'이 기타 문자 없이 한글만으로 쓰인 경우에는 사전의 처리를 따라 '십육강전/NNG'으로 분석한다. 기타 문자를 이용해서 표기했는지 한글만으로 표기했는지에 따라 달리 처리하게 되지만, 기타 문자는 분리하는 것이 원칙이라는 점을 고려한 것이다.

단, 숫자나 외국어로만 표기된 경우에는 모두 각각을 분석한다.

[예시] 6.25 [6/SN+./SP+25/SN]

(2) 명사 상당어의 분석

(가) 동사의 활용형이 따옴표 없이 문장 속에서 명사처럼 기능하는 경우는 원래 품사대로 분석한다.

[예시] 어디 가느냐가 그의 물음이었다. [가/VV+느냐/EF+가/JKS]

(나) 따옴표를 가진 성분이나 요소도 명사처럼 기능할 수 있으나, 원래 품사대로 분석한다.

[예시] 그것은 "는"이 아니라 "를"이다. ["/SS+는/JX+"/SS+이/JKC]

(다) 부사 뒤에 격조사가 쓰이는 것도 의미론적인 따옴의 효과에 의하여 부사가 명사적인 용법을 가지는 것이므로 분석은 '부사'로 한다.

[예시] 기름을 꽉 채우려면 가득을 누르세요. [가득/MAG+을/JKO]

나) 고유명사(NNP)

고유명사는 특정한 사물에 붙여진 이름으로, 기본적으로 최하의어에 속하는 대상을 서로 변별하기 위하여 붙인 이름이며, 원칙적으로 지시 대상만 가질 뿐 의미 내용은 가지지 않는다. 고유명사의 분석 기준은 매우 다양하므로, 본 지침에서는 아래에 제시하는 것만을 고유명사로 인정한다.

아래에 제시한 고유명사 부류에 속하는 말이 두 어절 이상에 걸쳐 나오는 경우가 있다. 이때에는 전체가 외국어로 구성된 말인지, 하나라도 고유어/한자어를 포함하고 있는지에 따라 달리 처리한다.

전체 어절이 외국어로 구성된 경우: 각 어절 모두를 NNP로 처리한다.
하나라도 고유어/한자어를 포함하고 있는 경우: 각각의 어절에 포함된 말이 무엇인지를 살펴 적절한 형태표지를 부여한다.

[예시] 블루 이즈 더 워미스트 컬러 (영화 제목)
[블루/NNP, 이즈/NNP, 더/NNP, 워미스트/NNP, 컬러/NNP]
[예시] 바람과 함께 사라지다 (영화 제목)
[바람/NNG+과/JKB, 함께/MAG, 사라지/VV+다/EF]

아래 지침은 주로 고유어/한자어를 포함하고 있는 고유명사 부류에 대한 설명임에 유의한다. 간혹 필요에 따라 전체가 외국어 구성인 경우에 대한 예시와 설명도 포함하였다.

(1) 인명, 종족명

(가) '씨(氏), 공(公), 군(君), 양(孃), 옹(翁), 대왕(大王)' 등 성 또는 이름 뒤에 같이 쓰이는 호칭어나 직책명은 분리해서 분석한다.

[예시] 남수/NNP+군/NNB, 김/NNP+씨/NNB, 최치원/NNP+옹/NNB, 케네디/NNP+씨/NNB
정/NNP+과장/NNG, 최/NNP+선생/NNG, 세종/NNP+대왕/NNG, 광개토/NNP+대왕/NNG

(나) 성과 이름, 호가 함께 쓰이면 하나의 단위로 분석한다.

[예시] 김철수/NNP, 이태백/NNP, 마르코폴로/NNP

(다) '씨, 군' 등과 달리 '가(哥)'는 접미사이므로, '김가(金哥), 이가(李哥)'는 파생어이다.

[예시] 김/NNP+가/XSN

(라) 사람 이름의 뒤에 '이'가 붙는 경우는 이름과 함께 하나의 단위로 분석한다.

[예시] 진현이/NNP+가/JKS

(마) 특정한 종족의 이름은 고유명사가 된다.

[예시] 알타이족/NNP, 피그미족/NNP, 돌궐족/NNP, 한족/NNP, 유대인/NNP

(바) 특정 동물에게 붙여진 이름도 인명에 준하여 고유명사가 된다.

[예시] 코코/NNP, 똘이/NNP

(사) 소설, 애니메이션 등 허구의 세계에서 쓰인 인명이나 동물명도 고유명사가 된다.

(2) 지명

<우리말샘>에 『지명』으로 올라 있는 단어 부류를 참고하여 지명 여부를 판단하고, 지명에 해당하는 부분과 지역의 종류를 나타내는 말을 묶어 전체를 고유명사로 처리한다.

(가) 내륙, 대륙, 지대, 주, 평원, 만, 늪, 습지, 분지, 사막, 유전, 탄전, 군락지
섬, 제도, 열도
바다, 해변, 포구, 강, 유역, 나루, 호수, 계곡, 연못, 갯벌, 폭포, 삼각주, 빙하, 피오르
길, 거리, 수로, 루트, 로드
산, 산맥, 산지, 화산, 동굴, 숲, 고개, 언덕, 오름, 구릉, 고원, 광산, 절리, 화산대 등의 이름

[예시] 카스피해/NNP, 템즈강/NNP, 태백산맥/NNP, 미시시피호/NNP, 갈라파고스제도/NNP, 갈론계곡/NNP, 감지해변/NNP, 강경포구/NNP, 강계분지/NNP, 강주연못/NNP, 거창분지/NNP, 고수동굴/NNP, 관동팔경/NNP, 그레이트빅토리아사막/NNP

(나) 도(道), 시(市), 읍(邑), 면(面), 리(里), 군(郡), 구(區), 동(洞), 골, 촌, 마을, 자치구, 연구단지, 관광단지, 공업지대, 지역, 지방, 지구 등의 이름은 그 구역의 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 서울특별시/NNP, 성북구/NNP, 강진군/NNP, 인창동/NNP, 빨래골/NNP, 해방촌/NNP, 고포마을/NNP, 네바다주/NNP, 경기지역/NNP, 경인공업지대/NNP, 관서공업지역/NNP, 광시좡족자치구/NNP, 가자지구/NNP, 서안지구/NNP

(3) 국가명 또는 왕조명

(가) 국가의 명칭, 또는 왕조의 명칭은 고유명사로 분석한다.

[예시] 대한민국/NNP, 조선/NNP, 코리아/NNP, 러시아연방/NNP, 미얀마연방공화국/NNP

(나) 다른 형태가 붙어 국가나 왕조의 존립 기간을 나타내는 경우 일반명사로 분석한다.

[예시] 대한제국기/NNG, 조선조/NNG

(다) '남한'과 '북한'을 의미하는 '남, 북, 남북'은 모두 일반명사와 고유명사를 구별한다. 남한을 뜻하는 '남'과 북한을 뜻하는 '북'을 고유명사로 분석한다.

[예시] 남/NNP+과/JC 북/NNP+의/JKG 의견/NNG 차이/NNG
남북/NNP 적십자/NNP+회담/NNG
[참고] 북미/NNP 회담/NNG
→ '북미' 자체는 <우리말샘> 등재어가 아니지만 구 표제어인 '북미^제네바^기본^합의서' 속에서 한 단어로 처리되고 있음을 참고하여 한 단어로 처리한다.

(라) 어떤 국가의 국민을 나타내는 '국가+인(人)'은 통합하여 일반명사로 분석한다.

[예시] 이집트인/NNG, 아제르바이젠인/NNG, 이스라엘인/NNG, 조선인/NNG

(마) 어떤 국가의 군대를 나타내는 '국가+군(軍)'은 통합하여 일반명사로 분석한다.

[예시] 미군/NNG, 북한군/NNG, 영국군/NNG, 일본군/NNG

(4) 건축물이나 시설물 혹은 구조물의 이름

<우리말샘>에 『지명』으로 올라 있는 단어 부류를 참고하여 고유명사 여부를 판단하고, 구조물명, 시설물명에 해당하는 부분과 구조물, 시설물의 종류를 나타내는 말을 묶어 전체를 고유명사로 처리한다.

(가) 도로, 항만, 항구, 터널, 대교, 철교, 뱃길, 운하, 댐, 공항, 터미널, 철도, 전철, 지하철 및 그 명칭과 함께 쓰이는 부대시설은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된

다.

[예시] 부산항/NNP, 대전역/NNP, 서울지하철/NNP, 인천공항/NNP, 테헤란로/NNP, 경부고속도로/NNP, 분당선/NNP, 경춘선/NNP, 정우터널/NNP, 강화대교/NNP, 경인아라뱃길/NNP

단, 어느 지역의 지하철에나 존재하는 '1호선, 2호선' 등은 특정성이 낮으므로 고유명사로 보지 않는다.

[예시] 1호선 [1/SN+호선/NNB]
→ 의존명사 '호'에 비분석 접사 '-선'이 결합한 구성으로, '-선'을 앞말에 붙여 '호선/NNB'로 처리한다.

(나) 해수욕장, 공원, 광장, 정원, 목장, 유원지, 유적지, 절터, 관광지, 테마파크, 전망대, 온천, 시장, 장터, 저수지, 기지, 묘지 등의 시설물도 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 속초해수욕장/NNP, 올림픽공원/NNP, 설악산국립공원/NNP, 한강시민공원/NNP, 광화문광장/NNP, 화정역광장/NNP, 황복사터/NNP, 가락시장/NNP, 노량진수산시장/NNP, 남극세종기지/NNP, 도라전망대/NNP, 현충원/NNP, 국립서울현충원/NNP

단, '농산물도매시장', '생활체육공원'과 같이 특정 시장이나 공원의 이름이 아니라 시장이나 공원의 유형을 나타내기 위해 쓰인 말은 고유명사로 보지 않는다.

'문화예술공원'은 서울특별시 서초구에 있는 특정 공원의 이름으로 쓰이기도 하고 공원의 유형을 나타내기 위한 말로 쓰이기도 하는데, 맥락을 구분하여 전자의 경우에는 고유명사로, 후자의 경우에는 일반명사로 분석한다.

[예시] 서초구 문화예술공원 (특정 공원의 이름일 때)
[문화예술공원/NNP]
우리 구에 문화예술공원을 설립합시다. (공원의 종류를 나타낼 때)
[문화/NNG+예술/NNG+공원/NNG]

(다) 배, 비행기와 같은 건조물의 이름은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다. [참고]와 같이 특정 집단의 우두머리 이름을 따 배에 빗대어 표현하는 경우에도 고유명사로 처리한다.

[예시] 최영함/NNP, 나로호/NNP
[참고] 신태용호/NNP

(라) 빌딩, 박물관, 극장 등 건물명은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다. 그리고 <우리말샘>에 구로 등재된 단위라고 하여도 그 통합형을 고유명사로 처리한다.

[예시] 서울역사/NNP, 세종문화회관/NNP, 개나리유치원/NNP, 청와대/NNP, 국회의사당/NNP
국립중앙박물관/NNP, 국립민속박물관/NNP, 구텐베르크박물관/NNP
신라호텔/NNP, 미도파백화점/NNP, 동궁예식장/NNP, 명보극장/NNP, 고대병원/NNP,
인천타워/NNP, 고마신사/NNP

(5) 회사, 상품, 학교, 정당, 기관이나 단체의 이름

(가) 특정 회사, 학교, 학회, 협회, 정당의 이름은 고유명사로 분석한다.

[예시] 삼성/NNP, 삼성그룹/NNP, 엘지전자/NNP, LG전자/NNP, 현대자동차서비스/NNP
코스닥/NNP, 나스닥/NNP, 코스피/NNP
고려대학교/NNP, 잠실고등학교/NNP, 송파중학교/NNP
국어학회/NNP, 영어영문학회/NNP, 한국사학회/NNP
대한축구협회/NNP, 승마협회/NNP, 프로야구선수협회/NNP
한나라당/NNP, 자유민주주의연합/NNP

회사, 학교, 학회, 협회, 정당의 이름이라고 해도 아래와 같이 한글 없이 기타 문자로만 표기된 경우에는 고유명사가 아니라 기호로 처리한다. 인명 등도 마찬가지이다.

[예시] LG에서 [LG/SL+에서/JKB]

(나) 특정 회사의 상품명과 브랜드명은 전체를 묶어 고유명사로 처리한다. 회사명과 상품명이 한 어절에 나왔을 때나 상품명과 상품의 종류를 나타내는 말이 한 어절에 나왔을 때에도 전체를 묶어 고유명사로 처리한다.

[예시] 초코하임/NNP, 코카콜라/NNP, 갤럭시S8/NNP, e편한세상/NNP, 쉐보레/NNP, 티맵/NNP
[예시] 농심새우깡/NNP, 칠성사이다/NNP, 신한SOL/NNP (회사명+상품명)
[예시] e편한세상아파트/NNP (브랜드명+종류명)
[예시] 하나에스라인적금/NNP (회사명+상품명+종류명)

(다) 상호명은 그 통합형을 고유명사로 처리한다. 상호명과 업종명이 한 어절에 나타난 경우에는 전체를 통합하여 고유명사로 분석하고, 상호명과 업종명이 두 어절로 분리되어 나타난 경우에는 상호명에 해당하는 부분을 통합하여 고유명사로 분석한다.

[예시] 한마음약국/NNP, 동일카센터/NNP
[예시] 다나은 이비인후과 [다나은/NNP, 이비인후과/NNG]

(라) 정부기관의 명칭 중 **지명 등의 고유명사를 포함하고 있어** 특정성이 높은 것만을 통합하여 고유명사로 처리한다. <우리말샘>에 구로 등재된 단위라고 하여도 그 통합형을 고유명사로 처리한다.

단, 특정 기관의 '지청, 지원, 지부', 특정 정당 소속의 '시당' 등은 그것이 지명과 함께 나타났더라도 모든 기관에 존재할 수 있어 특정성이 낮으므로 고유명사로 처리하지 않고 [예시]와 같이 단어별로 분리하여 일반명사로 분석한다.

하지만 '지방법원'의 준말인 '지법'은 지명과 함께 쓰일 경우 특정 기관을 가리키므로(예: 서울지법, 부산지법), '서울지법/NNP, 부산지법/NNP'로 처리한다.

고유명사를 포함하지 않아 고유명사에 들지 않는 정부기관 명칭은 단어별로 분리하여 처리한다.

[예시] 서울고등법원/NNP, 서울시경찰서/NNP, 서대문구치소/NNP
[예시] 여주/NNP+지청/NNG, 충북/NNP+지부/NNG, 서울시/NNP+당/NNG(서울시+당[명사])
[예시] 헌법/NNG+재판소/NNG, 국립/NNG+국어원/NNG, 여성/NNG+가족부/NNG

어떤 단어가 정부기관의 명칭인지, 건물명인지 혼동되는 경우가 있다. 이때에는 <우리말샘>의 뜻풀이를 참고하여, 주된 뜻풀이에 '기관'이라 되어 있으면 정부기관으로 판단하고, '건물'이라 되어 있으면 건물명으로 판단한다.

[예시] 청와대 (사전: 서울 경복궁 뒤 북악산 기슭에 있는 우리나라 대통령 **관저**)
→ 사전의 뜻풀이에서 건물명으로 처리되고 있으므로 [청와대/NNP]로 처리한다.
[예시] 경찰청 (사전: 안전 행정부 소속하에 설치되어 경찰 업무를 관장하는 정부 행정 **기관**)
→ 사전의 뜻풀이에서 정부기관명으로 처리되고 있으며 고유명사가 포함되어 있지 않으므로 [경찰청/NNG]로 처리한다.

(마) 연구소, 위원회, 협의회, 본부, 종교단체의 경우에는 **고유명사나 '국가, 전국, 국제, 세계' 및 이와 유사한 단어를 포함하고 있어** 특정성이 높은 것만을 통합하여 고유명사로 처리한

다.

고유명사에 들지 않는 연구소, 위원회, 협의회, 본부, 종교단체명은 단어별로 나누어 처리한다.

[예시] 한국전자통신연구소/NNP, 대한예수교장로회/NNP, 대한불교조계종/NNP, 서울자사고교장협의회/NNP, 한국야구위원회/NNP, 국가인권위원회/NNP, 국제통화기금/NNP
[예시] 조계종/NNP, 천태종/NNP (인명, 지명을 포함하고 있는 말임)
[참고] 감리교/NNG, 장로교/NNG (고유명사를 포함한 말이 아님)
[예시] 통일/NNG+연구소/NNG, 생활/NNG+체육/NNG+연구소/NNG,
통사/NNG+론/XSN+연구회/NNG, 입주자/NNG+대표/NNG+협의회/NNG,
수니파/NNG, 방송/NNG+통신/NNG+위원회/NNG
→ 이처럼 고유명사나 '국가, 전국, 국제, 세계'를 포함하지 않은 단체명은 단어별로 나누어 처리한다.

(바) 특정 부대, 스포츠팀, 그룹, 조직, 클럽에 붙여진 이름은 종류를 나타내는 말과 함께 묶어 고유명사로 처리한다. 이 밖에도 이 부류와 유사한 특정 집단에 고유하게 붙여진 이름을 고유명사로 처리한다.

[예시] 백마부대/NNP (부대명) cf) '육군', '해군', '공군', '해병대'는 일반명사임.
인디언스/NNP, 기아/NNP, 타이거즈/NNP, 서울FC/NNP (스포츠팀명)
트와이스/NNP, 송골매/NNP (그룹명)
서방파/NNP (폭력조직명)
위너스클럽/NNP (클럽명)
티파티/NNP (미국 정부의 건전한 재정 운용을 위해 세금 감시 운동을 펼치는 시민 중심의 신생 보수 단체.)

(사) 약어나 준말의 처리

고유명사가 축약된 형태(준말)로 쓰일 경우 본디말과 함께 준말도 인정하여 축약된 형태 그대로를 고유명사로 분석한다.

[예시] 육사/NNP, 고대/NNP, 자민련/NNP, 서울고법/NNP

(아) 단체장명의 처리

지명/단체명에 '-장(長)'과 같이 접미사가 결합하거나 '수(守)'와 같이 접사에 준하는 요소가 결합하여 단체장명이 만들어진 경우, 해당 요소를 바로 앞말에 붙여 분석한다. (가령

'헌법재판소'는 본 지침에서 '헌법/NNG+재판소/NNG'가 되므로, 여기에 '-장'이 결합한 경우 '헌법/NNG+재판소장/NNG'으로 처리한다.) 또한 단체장명은 일반명사임에 유의하여 형태 표지를 부여한다.

[예시]	가평군수(가평군/NNP+수)	[가평군수/NNG]
	서울시장(서울시/NNP+장)	[서울시장/NNG]
	헌법재판소장(헌법/NNG+재판소/NNG+장)	[헌법/NNG+재판소장/NNG]
	편찬위원회장(편찬/NNG+위원회/NNG+장)	[편찬/NNG+위원회장/NNG]
	영등포노인종합복지관장	[영등포노인종합복지관장/NNG]
	인사부장	[인사부장/NNG]
	상황실장	[상황실장/NNG]
	비대위원장(비대위/NNG+원+장)	[비대위원장/NNG]
	→ 이 단어는 '비대위'에 구성원을 나타내는 접사 '-원', 그리고 우두머리를 나타내는 접사 '-장'이 차례로 결합한 것으로 파악된다. 따라서 '비대위'에 '-원'이 결합하여 '비대위원'이 되고, 여기에 '-장'이 결합하여 '비대위원장'이 된 것으로 보아 전체를 하나의 명사로 분석한다. '영진위원장' 등도 마찬가지이다.	

접사가 아니라 '소장, 원장, 지사'와 같은 명사가 결합하여 단체장명이 만들어지는 경우가 있다. 그런 경우에는 아래와 같이 분석한다.

[예시]	헌법재판소소장	[헌법/NNG+재판소/NNG+소장/NNG]
	민족문화연구원원장	[민족/NNG+문화/NNG+연구원/NNG+원장/NNG]
	서울시의원	[서울시/NNP+의원/NNG]
	제주도지사	[제주/NNP+도지사/NNG]
	→ '제주도+지사'로 분석할 수도 있으나, 합성어 주의사항 ⑥에 따라 뒤쪽에 더 많은 음절수가 남는 '제주+도지사'를 선택함.	

(6) 책, 연극, 영화, 음악 등 창작물의 제목

창작물의 제목이 한 어절에 나타난 경우 통합하여 고유명사로 분석한다.

[예시] 삼국사기/NNP, 손자병법/NNP, 고래사냥/NNP
[예시] 동이/NNP, 봉오동전투/NNP(영화 제목), 신과함께/NNP, 토지/NNP

(7) 신문, 잡지, 방송 채널, 웹사이트 등 매체의 이름

매체명이 한 어절에 나타난 경우 통합하여 고유명사로 분석한다. 두 어절 이상으로 분리되어 나온 경우에는 각 어절에 맞는 형태표지를 부여한다. 단, 외국어로 구성된 매체명이 여러 어절로 나타났을 때는 아래 다)의 외국어 처리 지침에 따라 모든 어절을 고유명사로 처리한다.

[예시] 조선일보/NNP, 여성동아/NNP, 폭스뉴스/NNP, 티브이엔/NNP, 보람TV/NNP(개인 채널명)
[예시] 유튜브/NNP, 트위터/NNP, 네이버/NNP, 페이스북/NNP
[예시] 스포츠 서울 [스포츠/NNG, 서울/NNP]
→ 전체 외국어 구성이 아님: 분리된 각 단어에 적합한 형태표지를 부여함.
[예시] 뉴스 위클리 [뉴스/NNP, 위클리/NNP]
스포츠 투데이 [스포츠/NNP, 투데이/NNP]
뉴욕 타임즈 [뉴욕/NNP, 타임즈/NNP]
→ 전체 외국어 구성인 경우: 외국어 지침에 따름

(8) 언어명

언어명의 경우 '-어'의 형태만을 통합하여 고유명사로 인정한다. '한국말'과 같은 경우는 일반명사로 분석한다.

[예시] 한국어/NNP, 일본어/NNP, 영어/NNP, 알타이어/NNP, 네덜란드어/NNP

(9) 유물명, 식물명, 동물명

유물명과 아래와 같은 식물명, 동물명은 전체를 묶어 일반명사로 취급한다.

[예시] 청자상감국화무늬긴목병/NNG
북부점박이올빼미/NNG

(10) 위에서 명시되지 않은 부류는 모두 고유명사로 인정하지 않는다.

[예시] 임진왜란 (사건명) [임진왜란/NNG]
노벨평화상 [노벨/NNP+평화상/NNG]
네안데르탈인 [네안데르탈인/NNG]
무한도전 (특정 매체의 프로그램 이름) [무한/NNG+도전/NNG]

메이지 (연호)	[메이지/NNG]
고양국제꽃박람회 (행사명)	[고양/NNP+국제/NNG+꽃/NNG+박람회/NNG]
한국시리즈 (대회명)	[한국/NNP+시리즈/NNG]

다) 외국어의 처리

외국어는 아래의 방식에 따라 처리한다.

(1) <우리말샘>에 한 단어로 등재된 외국어

전체를 묶어 의미에 따라 NNG 또는 NNP로 처리한다. 그 외의 경우(<우리말샘>에 등재되지 않았거나 구로 등재된 것, <우리말샘>에 한 단어로 등재되었지만 원문에서 여러 어절로 분리되어 나타난 것)에 대한 처리는 아래의 (2)에 따른다.

[예시] 가든파티 (사전: 가든-파티)	[가든파티/NNG]
[예시] 마추픽추 (사전: 마추픽추)	[마추픽추/NNP]

(2) <우리말샘>에 등재되지 않았거나 구로 등재된 외국어

(가) 한 어절로 나타났든 두 어절 이상으로 나타났든 **전체 외국어 표현이 본 지침의 고유명사 부류에 든다면,** 그 고유명사를 이루는 **각 어절 모두를 (어절 내부 분석 없이) NNP로** 처리한다.

[예시] 레드제플린 (록 그룹 이름)	[레드제플린/NNP]
레드 제플린	[레드/NNP, 제플린/NNP]
[예시] 카미노데산티아고 (지명, 순례길 이름)	[카미노데산티아고/NNP]
카미노 데 산티아고	[카미노/NNP, 데/NNP, 산티아고/NNP]
[예시] 로버트다우니주니어 (인명)	[로버트다우니주니어/NNP]
로버트 다우니 주니어	[로버트/NNP, 다우니/NNP, 주니어/NNP]
페르디낭 드 소쉬르	[페르디낭/NNP, 드/NNP, 소쉬르/NNP]
cf) 루이9세	[루이/NNP+9/SN+세/NNB]
루이 9세	[루이/NNP, 9/SN+세/NNB]
[예시] 웻 앤 와일드 워터월드 (시설물명)	[웻/NNP, 앤/NNP, 와일드/NNP, 워터월드/NNP]
[예시] 블루 이즈 더 워미스트 컬러 (영화제목)	

[블루/NNP, 이즈/NNP, 더/NNP, 워미스트/NNP, 컬러/NNP]

(나) 위의 경우가 아니라면, 외국어를 포함하고 있는 각각의 어절에 대하여 다음과 같은 절차를 적용하여 처리한다.

(나-1) **각각의 어절 속에 포함된 외국어 요소가 한 단어라면,** <우리말샘> 등재 여부와 무관하게 그 단어를 의미에 따라 **고유명사 또는 일반명사로** 분석한다.

[예시] 마이너하다 [마이너/NNG+하/XSA+다/EF]

→ 미등재어인 '마이너'는 '마이너 감성', '메이저와 마이너의 경계' 등에서 볼 수 있듯이 다른 말과도 결합하여 쓰이므로 어근보다는 명사의 성격을 띠는 단어로 볼 수 있다. 외국어의 이런 특성을 고려하여, 한 단어에 해당하는 미등재 외국어를 어근(XR)이 아닌 체언류로 처리한다.

[예시] 마이너 리그 [마이너/NNG, 리그/NNG]

[예시] 라 리가 (축구 리그) [라/NNG, 리가/NNG]

(나-2) **한 어절 속에 포함된 외국어 요소가 둘 이상의 단어일 때에는** 그 외국어 요소를 **단어 단위로 분리한다.** 단어 단위를 판단할 때에는 해당 외국어에서의 표기법(띄어쓰기 여부)을 참고한다. 그 후 아래의 지침을 따른다.

[예시] 배팅글러브 [배팅, 글러브]

[예시] 아시안게임 [아시안, 게임]

[예시] 리우올림픽 [리우, 올림픽]

[예시] 보이그룹 [보이, 그룹]

[예시] 라리가 (축구 리그) [라, 리가]

(나-3) 분리된 각 단어가 본 지침의 고유명사 부류에 들거나 <우리말샘>에 단독 일반명사로 등재되어 있어서 **각각의 단어를 따로 처리할 수 있는 상황이라면, 각 단어를 분리하여 형태표지를 부여**한다.

[예시] 배팅글러브 ('배팅' 등재, '글러브' 등재) [배팅/NNG+글러브/NNG]

[예시] 아시안게임 ('아시안' 고유명사, '게임' 등재) [아시안/NNP+게임/NNG]

→ '아시안', '아메리칸', '브리티시' 등 고유명사의 형용사형을 모두 고유명사로 처리한다.

[예시] 리우올림픽 ('리우' 고유명사, '올림픽' 등재) [리우/NNP+올림픽/NNG]

(나-4) 분리된 각 단어 중 어느 하나라도 위의 방식에 따라 고유명사로 또는 일반명사로 **처리할 수 없다면,** 각 단어를 분리하지 않고 **전체를 묶어서 의미에 따라 고유명사 또는 일반명사로** 분석한다.

[예시] 보이, 그룹: '그룹'은 팀의 의미로 단독으로 등재되어 있으나, '보이'는 '소년'의 의미로서는 단독으로 명사로 등재되지 않음. '보이'의 처리가 어려우므로 전체를 묶어 일반명사로 분석함. [보이그룹/NNG]

[예시] 라, 리가: '라'와 '리가' 모두 고유명사 또는 일반명사로 처리하기 어려움. 대회명은 본 지침에서 고유명사에 들지 않으므로 전체를 묶어 일반명사로 분석함. [라리가/NNG]

(나-5) 위에 제시한 절차를 외국어를 포함하고 있는 모든 어절에 각각 적용한다. 예시는 다음과 같다.

[예시] 피지컬 트레이닝	[피지컬/NNG, 트레이닝/NNG]
[예시] 워터해저드	[워터해저드/NNG]
[예시] 골든 커리어 그랜드슬램 (기록명)	[골든/NNG, 커리어/NNG, 그랜드슬램/NNG]
[예시] 글로벌 북카페 (신문 코너명)	[글로벌/NNG, 북카페/NNG]
글로벌 북 카페	[글로벌/NNG, 북/NNG, 카페/NNG]

(다) 아래와 같이 **외국어의 '한 문장'이 한글로 전사되어 나타난 경우, 각 어절을 내부 분석 없이,** 그리고 각 단어의 <우리말샘> 등재 여부와 무관하게 **NA로** 처리한다.

[예시] 렛츠고	[렛츠고/NA]
[예시] 익스큐즈 미	[익스큐즈/NA, 미/NA]
[예시] 아이 러브 유	[아이/NA, 러브/NA, 유/NA]
[예시] 굿!	[굿/NA+!/SF]
[예시] 곤니치와	[곤니치와/NA]
[예시] 니하오	[니하오/NA]
[예시] 해피버스데이 투 유	[해피버스데이/NA, 투/NA, 유/NA]

라) 의존명사(NNB)

의존명사는 자립해서 쓰일 수 없는 명사로, 수식 성분을 반드시 동반해야 한다. 의존명사는 비단위성 의존명사와 단위성 의존명사로 나뉠 수 있으나, 본 분석에서는 이를 세분하지 않는다.

의존명사와 일반명사의 구분은 <우리말샘>에 따른다.

(1) 의존명사와 일반명사의 구분

(가) '연대, 연도'는 '년대, 년도'와 달리 일반명사이다.

[예시] 연도별로 정리된 자료	[연도/NNG]
몇 년도에 일어난 일	[년도/NNB]

(나) '월, 연, 일, 주, 달러, 원' 등은 독립되어 쓰일 경우 모두 일반명사의 자격을 가지므로 일반명사로 분석해야 한다.

[예시] 나는 월 30만원을 받는다.	[월/NNG]
달러의 가치는	[달러/NNG]
시간당 만원을 받는다.	[시간/NNG+당/XSN]

주의사항

'원'은 <우리말샘>에 의존명사로만 올라 있지만, 독립되어 쓰인 경우에는 일반명사로 분석한다. 다른 유사 경우도 이에 따라 처리한다.

[예시] 1달러를 원으로 환산하면	[원/NNG+으로/JKB]

골프에서 '2타를 줄였다' 등으로 쓰이는 '타'가 있는데, 이 말이 사전에 올라 있지 않다. 본 분석에서는 이와 같은 '타'를 일반명사로 분석한다.

[예시] 2타를 줄여	[2/SN+타/NNG+를/JKO]

(2) 단위를 나타내는 표현

(가) 길이, 무게, 수효, 시간 따위의 수량을 수치로 나타내는 단위들 중 '미터, 그램, 리터' 등은 의존명사(NNB)로, 외국어로 된 'm, g, l' 등은 기호(SW)로 분석한다.

(나) 일반명사가 단위적인 용법으로 쓰인 경우에는 의존명사가 아니므로 주의한다.

[예시]	사람, 그릇…	
	한 사람이 교실로 들어왔다.	[사람/NNG+이/JKS]
	자장면 한 그릇만 주세요.	[그릇/NNG+만/JX]

(3) '것'과 구어형 '거'의 분석

'거'의 형태를 그대로 인정하여 분석한다.

[예시]	공부할 거를 준비해 왔니?	[거/NNB+를/JKO]
	공부할 걸 가져왔니?	[거/NNB+ㄹ/JKO]
	연습할 건 있니?	[거/NNB+ㄴ/JX]
	먹을 게 모자라다.	[거/NNB+이/JKS]

2) 대명사(NP)

대명사는 그 자체로는 자신의 본유적 지시물을 가지지 않은 채, 다만 사람이나 사물 등 어떤 대상을 간접적으로 지시하는 품사이다. 단, 동일한 대명사가 방언이나 고어의 이형태를 가진 경우에는 이들도 대명사로 같이 분석한다.

가) 인칭 대명사

(1) 1인칭 대명사

[예시] 나, 내, 우리, 저, 제, 저희

(2) 2인칭 대명사

[예시] 너, 네, 그대, 당신, 댁

(3) 기타 대명사

[예시] 이이, 이분, 그이, 그분, 저이, 저분, 아무, 아무개, 누구, 무엇, 뭐, 어디, 언제, 자기, 걔, 쟤, 얘, 이것, 저것, 그것, 이거, 저거, 그거, 여기, 저기, 거기, 이곳, 그곳, 저곳, 어디, 모(某), 모모(某某)

나) 대명사와 관형사의 두 가지 분석이 가능한 단어

(1) '모(某)'는 관형사와 대명사로 분석될 수 있으므로 주의를 요한다.

[예시]	모 기업체	[모/MMD]
	김 모씨	[모/NP+씨/NNB]

(2) '모모(某某)'도 위와 같이 분석될 수 있다.

[예시]	모모가 말했다	[모모/NP+가/JKS]
	모모 기관의 조사를 마쳤다	[모모/MMD]

다) 대명사의 이형태 분석

(1) '이것, 그것, 저것; 이거, 그거, 저거'는 분석하지 않고 대명사로 인정한다. '~거'의 경우, '~거'의 형태를 그대로 인정하여 분석한다.

[예시]	난 저거를 먹을래.	[저거/NP+를/JKO]
	나는 여태 그걸 믿어 왔단다.	[그거/NP+ㄹ/JKO]

(2) 다음과 같이 원형을 밝힐 수 있는 대명사는 원형대로 분석한다.

[예시]	내	이제부터는 내 명령을 따라라.	[나/NP+의/JKG]
	내게	내게 전자우편으로 알려 다오.	[나/NP+에게/JKB]
	네게	어제 네게 보낸 선물이 잘못되었다.	[너/NP+에게/JKB]
	제게	문제가 있다면 제게 말씀해 주세요.	[저/NP+에게/JKB]
	누가	누가 전화를 하는지 보고해라.	[누구/NP+가/JKS]
	뉘	뉘 집 애기가 이렇게 울고 있는 거야?	[누구/NP+의/JKG]
	뭐가	도대체 뭐가 문제라는 거야?	[뭐/NP+가/JKS]

(3) '제'의 경우, '제/NP+가/JKS'를 제외하고는 모두 '저/NP+의/JKG'로 분석한다.

[예시]	제가 갈 것입니다.	[제/NP+가/JKS]
	철수는 제 잘못을 안다.	[저/NP+의/JKG]
	제 무게를 못 견디다.	[저/NP+의/JKG]

3) 수사(NR)

수사는 사물의 수량이나 차례를 나타내는 품사를 말한다.

가) 수사의 종류

(1) 양수사

[예시] 하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉, 열, 스물, 서른, 마흔, 쉰, 예순, 일흔, 여든, 아흔, 백
한둘, 두서넛, 서넛, 너덧, 네다섯, 네댓, 대여섯, 예닐곱, 일여덟, 일고여덟, 열두서넛, 열대여섯, 열일고여덟, 스물두서넛
일, 이, 삼, 사, 오, 육, 칠, 팔 구, 십, 백, 천, 만, 억, 조
기십, 기백, 기천,
수십, 수백, 수천, 수만, 수억, 수십만, 수백만, 수천만

(2) 서수사

[예시] 첫째, 둘째, 셋째, 넷째, …, 열째, 열한째,…, 스물한째,…,
아흔아홉째, 백째, 백한째, …

주의사항

① 값이 없는 수를 나타내는 '영', '공'은 사전에 명사로 올라 있으나 수사로 분석한다.

② 외국어로 된 말인 '제로, 원, 투, 쓰리' 등은 우리말 단위 명사와 잘 결합하지 않는 등 일반적인 수사와 다른 성격을 지니므로 일반명사로 분석한다.

③ 복수의 수사가 한 어절 내에 나타날 때에는 전체를 통합해서 분석한다.

[예시] 백만오천삼십사 [백만오천삼십사/NR]
공일공일이삼사오륙칠팔 [공일공일이삼사오륙칠팔/NR]
[참고] '일일구', '일일이' 등 <우리말샘>에 명사로 올라 있는 것은 명사로 분석한다.

④ ‘하나’는 <우리말샘>에 그 품사가 명사와 수사로 되어 있지만 본 지침에서는 수사로 분석한다.

[예시]	광에 가서 물건 하나만 가져오렴.	[하나/NR+만/JX]
	우리는 하나로 뭉쳤다.	[하나/NR+로/JKB]

⑤ 때로 수사와 수관형사의 구별이 애매한 경우가 있다. 이 분석에서는 **다음과 같이 특이한 형식을 가진 예만을 수관형사로 취급하고**, 그 밖의 것들은 모두 수사로 분석한다. 사전에 수사와 관형사 동일 형태로 등재된 ‘몇’과 ‘몇몇’ 역시 모두 수사로 분석한다. 이는 <우리말샘>의 품사 처리와는 다른 방식임에 유의한다.

[예시]	한, 한두, 한두어, 두, 두어, 두세, 두서너, 세, 석, 서, 서너, 네, 너, 넉

⑥ 순서를 나타내는 ‘제일, 제이’ 등은 접두사 ‘제-’와 수사의 결합으로 분석한다.

[예시]	제일, 제이, 제삼, 제사, 제오 … 제구십구, 제백…	[제/XPN+일/NR], [제/XPN+이/NR], …

⑦ 순서를 나타내는 ‘첫째, 둘째’ 등에 포함된 접미사 ‘-째’는 분석하지 않는다. ‘첫 번째’, ‘두 번째’ 등 의존명사 ‘번째’에 포함된 접미사 ‘-째’는 분석한다.

[예시]	첫째, 둘째, 셋째, 넷째, 다섯째, …, 아흔아홉째, …	[첫째/NR], [둘째/NR], …
[예시]	첫 번째	[번/NNB+째/XSN]

나 용언

용언은 동사, 형용사, 지정사를 가리킨다. 용언 범주에서는 분석 대상이 본용언일 경우에만 동사와 형용사로 구분하여 표시하고, 보조용언의 경우에는 보조동사와 보조형용사를 구분하지 않고 ‘VX’라는 하나의 표지만을 준다. 또한 학교 문법에서 서술격조사로 다루는 ‘이다’는 조사의 범주에 넣지 않고 ‘지정사’라는 용언의 하위범주에 넣기로 한다. 지정사는 다시 긍정 지정사(VCP)와 부정 지정사(VCN)로 세분된다.

1) 동사(VV)

동사는 사물의 움직임이나 작용을 나타내는 용언을 말한다. 동사는 일반적으로 목적어의 필요성 여부에 따라 자동사, 타동사로 나누기도 하지만, 본 분석에서는 그것을 위한 별도의 표지를 세분하지 않고 모두 'VV'로 표시한다.

주의사항

'있다'는 동사 용법과 형용사 용법을 모두 가지고 있다. '있다'는 대개의 경우 형용사로 쓰이는 것으로 보아, '있다'가 동사로 쓰였다는 적극적인 증거가 있을 때에만 동사로 분석하고 나머지 경우는 형용사로 분석한다.

'-고 있다', '-어 있다'형으로 쓰여 앞의 사태가 진행/지속되고 있음을 나타내거나 앞의 사태가 끝나고 그 결과가 유지되고 있음을 나타낸다면 그때의 '있다'는 보조용언(VX)임에 유의한다.

형용사 '있다'의 특징

① '존재하다'의 뜻을 갖는 것은 형용사이다.

[예시] 신이 있다 / 날지 못하는 새도 있다 / 기회가 있다 / 증거가 있다
짜임새 있다 / 쓸모 있다 / 진정성 있다 / 경쟁력 있다 / 가능성 있다 / 필요 있다

② '수 있다', '바 있다', '적이 있다' 구성의 '있다'도 모두 형용사이다.

③ 종결어미 '-다'와 바로 결합하여 '있다.'형으로 쓰이면 형용사이다.

[예시] 이런 경우도 있다. / 그는 서울에 있다.

④ '~에 있어서' 구성의 '있다'도 형용사이다.

[예시] 인간에게 있어서 중요한 것은 사랑이다.

⑤ '누가 어떤 자격으로 있다' 구성의 '있다'도 형용사이다.

[예시] 그는 지금 대기업의 과장으로 있다.
그는 대기 선수로 있다가 출전권을 얻었다.

⑥ 내포문에서 '있다'가 쓰였을 때에는, 종결형으로 바꾸었을 때 '있다.'를 사용하여 표현할 수 있는 경우면 모두 형용사로 판단한다.

[예시] 서울광장에서 있었던 콘서트가 그 예이다.
→ '서울광장에서 콘서트가 있다.'로 바꾸어도 문장이 성립하므로, 형용사로 판단한다.
친구와 둘만 있는 상황이 되면...
→ '(나는 지금) 친구와 둘만 있다.'로 바꾸어도 문장이 성립하므로 형용사로 판단한다.

※ '머물다'의 뜻을 갖는다고 해서 모두 동사로 판단하지 않는다. '머물다'의 의미는 아래와 같이 사전에 동사로도, 형용사로도 기술되어 있다. 따라서 위 ⑥에서 언급했듯이 '있다.'형으로 바꾸어도 문장이 성립되면 '머물다'의 뜻이어도 모두 형용사로 판단하기로 한다. '있다.'형으로 바꿀 수 없거나 '-는다', '-어라', '-자'처럼 동사와 결합하는 어미와 함께 나타났을 때에만 동사로 판단한다.

있다1 [I] 동사
「1」 사람이나 동물이 어느 곳에서 떠나거나 벗어나지 아니하고 머물다.
예) 그는 내일 집에 있는다고 했다.
있다1 [II] 형용사
「2」 사람이나 동물이 어느 곳에 머무르거나 사는 상태이다.
예) 그는 한동안 이 집에 있었다.

⑦ '있다'가 기간을 나타내는 부사어와 함께 쓰일 때에는 종결형 '있다.'를 사용하여 표현하는 것이 어색하다. 이런 경우 동사로 판단한다.

[예시] 그는 노쇠해서 이 자리에 오래 있기 힘들다.
→ '그는 이 자리에 오래 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.
1시간가량 조용히 있다가 갑자기 일어나 총을 꺼내 들었다.
→ '그는 1시간가량 조용히 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.
오늘은 덕수궁 지하도에 더 있다 갈게요.
→ '그는 덕수궁 지하도에 더 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.

동사 '있다'의 특징

① '-는다'(평서), '-어라', '-자', '-읍시다' 등(명령, 청유)이 결합한 것은 동사이다.
② '얼마의 시간이 경과하다'의 뜻일 때에는 동사이다.

[예시] 10분 있다 만나자. / 얼마 안 있어 기다리던 시간이 왔다.

③ 아래 예시와 같이 '잘'과 결합한 '있다'는 동사로 판단한다. '잘 계세요'로 치환이 가능하다는 점에서 동사의 행태를 보이기 때문이다. 단, '내 그물이 잘 있나.'에서처럼 주어가 사람이 아

닌 경우에는 동사로 판단하지 않는다.

[예시] (헤어질 때) 잘 있어요. / 아버지는 잘 있느냐. / 잘 있었니.
cf) 잘 계세요. / 아버지는 잘 계시느냐. / 잘 계셨습니까.

④ '-고 싶다', '-려(고) 하다'는 주로 동사와 결합하므로, 이와 결합한 '있다'는 동사로 판단한다.

[예시] 나도 그 자리에 있고 싶다.
나도 여기 있으려고 한다.

⑤ '가만히 있-', '마냥 있-'은, '가만있다'가 동사임을 참고하여, 또 '철수가 가만히 있다.', '철수가 마냥 있다.' 같은 표현이 빈번히 쓰이지 않는 것을 고려하여 동사로 판단한다.
단, 종결어미 '-다'가 바로 결합하여 '가만히 있다.', '마냥 있다.'로 쓰였다면 형용사로 판단한다.

⑥ '있은 지', '있은 후' 등에서 나타나는 '있은'의 '있-'은 동사로 판단한다. '-은'이 결합하여 과거를 나타내는 것이 동사의 특성이기도 하고, '-ㄴ 지', '-ㄴ 후' 등도 주로 동사와 결합하여 쓰이기 때문이다.

[예시] 그 일이 있은 지 수일이 지났다.

2) 형용사(VA)

형용사는 사물의 성질이나 상태를 나타내는 용언을 가리킨다.

주의사항

① 사전에 형용사로 등재된 단어가 동사와 같은 활용을 보일 때가 있다. 그러나 그럴 때에도 사전을 따라 형용사 형태표지를 부여한다.

[예시] 현실과 동떨어지는 문제가 있다. [동떨어지/VA+는/ETM]
→ '동떨어지다'는 <우리말샘>에 형용사로 등재되어 있다. 위의 예에서는 관형형 어미 '-는'과 결합하여 동사와 같은 활용 양상을 보여 주고 있으나, 본 지침에서는 이러한 경우에도 <우리말샘>의 품사를 따라 형용사로 분석한다.

② '못하다'는 <우리말샘>에 보조용언, 형용사, 동사 모두로 등재되어 있다. 따라서 용법에 맞

게 품사를 구별하여 분석해야 한다.

[예시] 노래를 못한다.	[못/MAG+하/XSV+ㄴ다/EF+./SF]
[예시] 맛이 예전만 못하다.	[못/MAG+하/XSA+다/EF+./SF]
못해도 열 명은 올 것이다.	[못/MAG+하/XSA+아도/EC]
[예시] 밥을 먹지 못한다.	[못하/VX+ㄴ다/EF+./SF]
옳지 못하다.	[못하/VX+다/EF+./SF]
보다 못해 간섭을 했다.	[못하/VX+아/EC]

3) 보조용언(VX)

이 분석에서는 보조용언을 보조동사와 보조형용사로 하위 구분하지 않는다.

가) 보조용언 분석 원칙

(1) 보조용언의 후보는 <우리말샘>에 그 쓰임이 제시되어 있어야 한다.

(2) 보조용언 앞에는 반드시 다른 용언이 위치해 있어야 한다.

(3) 보조용언이 동시에 두 개 이상이 연결되어 나타날 수도 있다.

(4) 본용언과 보조용언의 결합형이 <우리말샘>에 하나의 어휘로 등재되어 있으면 보조용언을 따로 분석하지 않고 전체를 하나의 용언으로 처리한다. 특히 '-어하다', '-어지다' 결합형이 사전에 하나의 어휘로 올라 있는 경우가 많으므로 유의한다.

[예시] 아이를 예뻐하고	[예뻐하/VV+고/EC]
[예시] 눈이 동그래졌다	[동그래지/VV+었/EP+다/EF]

나) 보조용언의 예

보조용언의 예시는 다음과 같다. 이 목록은 <우리말샘>을 참고한 것이다.

가다	책을 다 읽어 간다.	[가/VX+ㄴ다/EF+./SF]
가지다	일을 그렇게 해 가지고는 기일을 맞출 수 없다.	[가지/VX+고는/EC]
계시다	손님께서 와 계십니다.	[계시/VX+ㅂ니다/EF+./SF]
나가다	정책을 추진해 나가는 과정에서 문제가 생겼다.	[나가/VX+는/ETM]
나다	일을 마치고 나니 상쾌하다.	[나/VX+니/EC]
내다	힘들겠지만 잘 견뎌 내야 한다.	[내/VX+아야/EC]
놓다	약속을 잡아 놓고 출장을 가다니	[놓/VX+고/EC]
달다	이번 시험 문제의 정답을 알려 다오.	[달/VX+오/EF+./SF]
대다	자꾸 졸라 대는 통에 그만 허락해 주고 말았다.	[대/VX+는/ETM]
두다	남겨 둔 돈도 이제 바닥이 났다.	[두/VX+ㄴ/ETM]
드리다	염려를 끼쳐 드려 송구하옵니다.	[드리/VX+어/EC]
들다	도무지 내 말은 믿으려 들지 않는다.	[들/VX+지/EC]
말다	어렵더라도 희망을 잃지 말아야 한다.	[말/VX+아야/EC]
먹다	나는 오늘도 약속을 잊어 먹었다.	[먹/VX+었/EP+다/EF+./SF]
못하다	그 참상을 차마 보지는 못할 것이다.	[못하/VX+ㄹ/ETM]
버리다	음식이 다 타 버렸다.	[버리/VX+었/EP+다/EF+./SF]
보다	이제는 새벽이 오는가 보다.	[보/VX+다/EF+./SF]
빠지다	썩어 빠진 생선을 사오다니	[빠지/VX+ㄴ/ETM]
싶다	너를 보고 싶다.	[싶/VX+다/EF+./SF]
쌓다	꼬치꼬치 물어 쌓는 통에 정신이 없었다.	[쌓/VX+는/ETM]
아니하다	일이 순리대로 풀리지 아니했다.	[아니하/VX+았/EP+다/EF+./SF]
않다	시간이 지나도 기차는 오지 않았다.	[않/VX+았/EP+다/EF+./SF]
오다	날이 밝아 온다.	[오/VX+ㄴ다/EF+./SF]
있다	그녀는 검정 옷을 입고 있었다.	[있/VX+었/EP+다/EF+./SF]
주다	아버지는 아기에게 동화책을 읽어 주었다.	[주/VX+었/EP+다/EF+./SF]
지다	평소보다 깨끗해진 내 방이 너무 좋다.	[깨끗하/VA+아/EC+지/VX+ㄴ/ETM]
치우다	다섯 명이 10인분의 식사를 먹어 치웠다.	[치우/VX+었/EP+다/EF+./SF]
터지다	끓인 지 오래 되어서 라면이 불어 터졌다.	[터지/VX+었/EP+다/EF+./SF]
하다	나귀를 쉬게 하는 것이 좋겠다.	[하/VX+는/ETM]

주의사항

① 다음과 같은 어절은 <우리말샘>에서 보조용언으로 취급되고 있으나, 여기서는 '의존명사+접사' 또는 '의존명사+보조용언'으로 분석한다. 이들 앞에는 항상 관형어가 온다는 분포적인 특성을 중시한 것이다.

[예시] 양하다/체하다/척하다/듯하다/법하다/뻔하다	[양/NNB+하/XSV+다/EF]
듯싶다	[듯/NNB+싶/VX+다/EF]

② <우리말샘>에서 보조용언으로 취급되는 '버릇하다'의 경우, 일반명사 '버릇'과 크게 구별되지 않으므로 '버릇'은 명사로 분석한다.

[예시] 자꾸 울어 버릇한다.	[버릇/NNG+하/XSV+ㄴ다/EF+./SF]

4) 지정사(VC)

지정사는 학교 문법의 서술격 조사에 대응되는 것인데, 용언과 같이 활용한다는 특성을 중시한 술어이다. 여기서는 학교 문법의 '이다'를 긍정 지정사로, '아니다'를 부정 지정사로 하위 구분한다. 일반적으로 '아니다'는 형용사로 다루어지기도 하나, 여기서는 '아니다'가 '이다'의 부정형이라는 점을 중시하여 '부정지정사'로 다룬다.

[예시] 철수는 매우 우수한 학생이다.	[학생/NNG+이/VCP+다/EF+./SF]
철수는 모범적인 학생이 아니다.	[아니/VCN+다/EF+./SF]

가) 지정사 '이/VCP'를 복원해야 하는 경우

(1) 체언에 어미가 직접 연결된 경우

[예시] 철수는 훌륭한 교사다.	[교사/NNG+이/VCP+다/EF+./SF]

(2) 조사에 어미가 직접 연결된 경우

[예시] 우리가 그를 본 것은 서울에서다.	[서울/NNP+에서/JKB+이/VCP+다/EF+./SF]

(3) '-였다'

[예시] 그 당시 나는 아이였다.	[아이/NNG+이/VCP+었/EP+다/EF+./SF]

(4) 어미 '-라고, -라는, -라도, -라며, -라면서, -라서'

[예시] 나는 그에게 절교라고 말했다.	[절교/NNG+이/VCP+라고/EC]
나는 친구라는 말이 좋다.	[친구/NNG+이/VCP+라는/ETM]
"집에 간다"라는 말에 놀랐다.	[가/VV+ㄴ다/EF+"/SS+이/VCP+라는/ETM]
→ '"집에 갈걸"이라는 말'을 [가/VV+ㄹ걸/EF+"/SS+이/VCP+라는/ETM]으로 분석하게 됨을 참고하여, '"집에 간다"라는 말'에서도 '이/VCP'를 복원한다. 다른 유사 경우도 마찬가지로 처리한다.	
집에 간다라는 말에 놀랐다.	[가/VV+ㄴ다/EF+이/VCP+라는/ETM]
나이가 어린 자라도 존중해 주어야 한다.	[자/NNB+이/VCP+라도/EC]
그는 최고라며 나를 추켜 주었다.	[최고/NNG+이/VCP+라며/EC]
"바보"라며 놀렸다.	["/SS+바보/NNG+"/SS+이/VCP+라며/EC]
그는 실수라면서 얼버무렸다.	[실수/NNG+이/VCP+라면서/EC]
"밥을 먹고"라면서 화를 냈다.	[먹/VV+고/EC+"/SS+이/VCP+라면서/EC]
너는 부자라서 우릴 이해하지 못할 것이다.	[부자/NNG+이/VCP+라서/EC]

(5) 아래와 같이 인용문 뒤에서 '하-'가 생략된 채 쓰인 '-며', '-는' 등은 '하-'의 복원 없이 형태 표지를 부여한다.

[예시] 얼마나 친절하냐?"며	[친절/NNG+하/XSA+냐/EF+?/SF+"/SS+며/EC]
얼마나 친절하냐?"는	[친절/NNG+하/XSA+냐/EF+?/SF+"/SS+는/ETM]

다 수식언

1) 관형사(MM)

관형사는 체언 앞에서 그것을 꾸미는 품사를 말한다. 관형사는 지시관형사(MMD), 수관형사(MMN), 성상관형사(MMA)로 세분하여 분석한다.

[예시]	한	한 가정	[한/MMN]
	그까짓	그까짓 일	[그까짓/MMD]
	그	그 문제	[그/MMD]
	이	이 사람	[이/MMD]

가) 지시관형사(MMD)

'이, 그, 저'와 같이 발화 현장이나 문장 밖에 존재하는 대상을 가리키는 관형사를 지시관형사로 분석한다. '어느, 무슨, 웬'과 같이 정해지지 않은 것을 대신하는 관형사, '이내 신세'의 '이내'와 같이 인칭 의미를 나타내는 관형사도 지시관형사로 분석한다. 이 밖에 '귀(貴)'와 '본(本)'은 청자 측과 화자 측을 지시하고 '동(同)'은 공간을, '현(現)'과 '전(前)'은 시간을 지시한다는 점에서 지시관형사로 볼 수 있다.

[예시] 이, 그, 저, 요, 고, 조, 이런, 그런, 저런, 다른, 타(他), 어느, 무슨, 웬, 어떤, 아무, 아무런, 귀(貴), 본(本), 동(同), 현(現), 전(前), 모(某), 그까짓, 각(各), 매(每), 오른, 왼

나) 수관형사(MMN)

(1) 체언 앞에서 사물의 수량이나 차례를 나타내는 관형사를 수관형사로 분석한다. 단 '다섯, 여섯' 등 수사와 수관형사의 형태가 동일한 경우에는 수사로 분석한다.

[예시] 한, 두, 세/서/석, 네/너/넉, 닷, 엿, 스무, 한두, 두세, 서너, 두서너, 일이, 이삼, 삼사, 여러, 모든, 온, 온갖, 갖은, 전(全), 첫, 양(兩)

(2) 복수의 수사와 수관형사가 한 어절 내에 나타날 때에는 전체를 통합해서 수관형사로 분석한다.

[예시]	스물한	[스물한/MMN]
	십수	[십수/MMN]

(3) '한'은 다음과 같이 수관형사 또는 성상관형사로 분석한다.

[예시]	책 한 권	[한/MMN] (‘하나’의 뜻)
	한 마을에 효자가 살고 있었다.	[한/MMN] (‘어떤’의 뜻)
	동생과 나는 한 이불을 덮고 잔다.	[한/MMN] (‘같은’의 뜻)
	한 20분쯤 걸었다.	[한/MMA] (‘대략’의 뜻)

다) 성상관형사(MMA)

체언의 성질이나 상태를 나타내는 관형사를 성상관형사로 분석한다.

[예시] 새, 헌, 옛, 순(純), 구(舊), 주(主), 약(約), 양대(兩大), 만(滿) 10세, 단(單), 총(總)

주의사항

① ‘지시, 성상, 수’ 중 어느 한 쪽으로 보기 힘든 관형사는 모두 ‘성상 관형사’의 테두리에 포함시킨다.

② 관형사는 때로 문맥에 따라 다른 품사로 분석될 가능성이 있으니 문맥을 잘 살펴서 분석해야 한다.

[예시]	관형사, 명사 통용	
	전 학기에 장학금을 받았다.	[전/MMD]
	그 사람을 전에 본 적이 있다.	[전/NNG]
[예시]	관형사, 부사 통용	
	단 세 명이서 그 일을 꾸몄다.	[단/MMA]
	단, 그 일은 해서는 안 된다.	[단/MAJ]
[예시]	관형사, 명사, 부사 통용	
	이내 마음을 어찌 알리요.	[이내/MMD]
	아침 들판에 이내가 끼었다.	[이내/NNG]
	그는 이내 떠나갔다.	[이내/MAG]

③ 수사가 명사를 단독으로 수식하는 경우 그것을 관형사로 분석하기 쉬우나, ‘수’를 나타내는 말 가운데서 앞서 언급한 수관형사를 제외하고는 수사는 오로지 수사로만 분석한다. 즉, 수사의 관형사적 쓰임을 인정하지 않는 것이다. 따라서 다음과 같이 ‘다섯’은 모든 환경에서

중의성 없이 '수사'로만 분석된다.

[예시] 다섯이 먹기에 충분하다.	[다섯/NR+이/JKS]
다섯 명이 앉아 있었다.	[다섯/NR]

2) 부사(MA)

부사는 주로 용언을 꾸며서 그 뜻을 더 세밀하고 분명하게 해 주는 품사를 말한다. 여기서는 부사를 세분하지 않고, 접속부사와 일반부사로만 나누기로 한다.

가) 접속부사(MAJ)

<우리말샘>에 등재된 접속부사만을 대상으로 접속부사 표지를 부여한다.

주의사항

① 접속부사는 종종 용언의 활용형으로도 쓰일 수 있으므로 주의한다.

[예시] 그래서 마지막에는 조심하라고 했지?	[그래서/MAJ]
상황이 그래서 영희가 결석을 했구나.	[그렇/VA+어서/EC]

② '그리고나서'의 분석

[예시] 그리고나서	[그리/MAG+하/XSV+고/EC+나/VX+아서/EC]

③ '그래도'는 용언의 활용형일 수도 있고 접속부사일 수도 있다. 두 용법을 구별하여 표지를 부여해야 하며, 용언의 활용형일 때는 아래와 같이 분석한다. 동사 '그러다'의 활용형인지 형용사 '그렇다'의 활용형인지가 불분명하고 두 가지 해석이 모두 가능할 때는 형용사 '그렇다'의 활용형으로 판단한다.

[예시] (누가) 그래도	[그러/VV+어도/EC]
(상황이) 그래도	[그렇/VA+어도/EC]

④ ‘그런데도’는 [그렇/VA+ㄴ데/EC+도/JX]로 분석한다.

나) 일반부사(MAG)

주의사항

① 일반부사는 종종 일반명사와 동일형태를 띠고 있어 구분이 어려운 경우가 있다. 이들은 뒤에 조사가 결합하느냐의 여부와, 문맥에서 후행 명사를 수식하느냐의 여부에 따라 부사와 명사로 분석될 수 있다.

[예시]	너의 진짜 속셈이 무엇인지 말해 봐라.	[진짜/NNG]
	그 수학 문제는 진짜 어려웠다.	[진짜/MAG]
	지금이 공부하기 딱 좋은 때이다.	[지금/NNG+이/JKS]
	나는 지금 막 집에 도착했다.	[지금/MAG]

② 부사적인 용법을 가졌음에도 불구하고 일반부사가 아닌 일반명사로만 <우리말샘>에 등재되어 있는 단어는 오로지 일반명사로만 분석한다.

[예시] 구석구석, 여기저기, 오랫동안, 이곳저곳, 좌우간, 처음, 최근

③ 일반부사로 분석하기 쉬운 활용상의 불완전동사인 ‘덩달아, 더불어’는 모두 동사로 옳게 분석해야 함에 주의한다.

[예시]	너는 덩달아 왜 난리니?	[덩달/VV+아/EC]
	우리 함께 더불어 살아가자.	[더불/VV+어/EC]

④ ‘명사+없이’는 원칙적으로 ‘일반명사+없이/MAG’로 태깅하지만, 아래와 같이 하나의 단어로 굳어져 사전에 등재된 경우는 ‘없이’ 통합형 자체를 하나의 일반부사로 분석한다.

[예시] 관계없이, 그지없이, 꾸밈없이, 끊임없이, 난데없이, 남김없이 등

라 독립언

1) 감탄사(IC)

감탄사는 화자의 부름이나 느낌, 놀람이나 대답을 직접적으로 나타내는 품사를 말한다.

[예시] 그럼, 야호, 어머, 앗, 아, 예, 그래, 아니(요), 글쎄, 참, 아이구, 와아, 오호, 세상에

주의사항

① 사람이 입으로 직접 내는 소리를 대상으로 하되, 흉내를 내는 의도가 없는 것과 본능적인 놀람이나 느낌을 나타내는 것을 대상으로 한다. 또한 감탄사와 혼동되는 부사로서 음성상징 어류의 부사어가 있는데, 이는 감탄사가 아닌 일반부사로 분석한다.

[예시] 야호! 드디어 정상이다. [야호/IC+!/SF]
쿨럭쿨럭 기침을 했다. [쿨럭쿨럭/MAG]

② 동물의 울음소리 등은 감탄사가 아니라 일반부사로 분석한다.

[예시] 검둥이는 멍멍 짖으며 수풀 속으로 뛰어들어갔다. [멍멍/MAG]

③ 욕이나 욕설을 나타내는 말은 전체를 감탄사로 분석한다.

[예시] 빌어먹을! [빌어먹을/IC+!/SF]

④ '뭐'는 문맥에 따라 대명사와 감탄사의 두 가지 쓰임이 있다.

[예시] 뭔지도 모른 채 [뭐/NP+이/VCP+ㄴ지/EF+도/JX]
신문에 뭐 대단한 특종이라도 실렸습니까? [뭐/IC]

⑤ 한 어절이 비정상적으로 늘어나거나 비정상적으로 늘어난 것에 다른 기호가 개입되었을 경우 분석불능 범주(NA)로 분석한다.

[예시] 그러어엄/NA, 으~어~이/NA

⑥ 구어에서 나타나는 담화 표지는 <우리말샘>을 참고로 하여 감탄사 표지(IC)를 부여한다. 물결표는 분석하지 않는다.

[예시] 저~, 음~, 저기~	[저/IC, 음/IC, 저기/IC]
어~, 그~	[어/IC, 그/IC]

마 관계언

조사는 주로 체언과 결합하여 다른 말과의 문법적 관계를 나타내거나, 특별한 뜻을 더해 주는 품사를 말한다. 조사는 그 수효가 많으므로 본 지침에서는 일부 사례만을 제시하였으며, 조사의 전체 목록은 <우리말샘>을 따르는 것을 원칙으로 한다. 조사는 크게 격조사, 보조사, 접속조사로 나뉘는데, 그 구분 역시 <우리말샘>을 따른다.

주의사항

한국어는 조사가 여러 개 결합하는 경우가 많은데, 조사 결합형은 아래와 같은 방식으로 세분 여부를 결정한다.

① 조사 결합형이 <우리말샘>에 등재되어 있지 않으면 각 조사를 분리하여 분석한다.

[예시] 부산에서도 대형 사고가 있었다.	[부산/NNP+에서/JKB+도/JX] (‘에서도’ 미등재)
그녀와의 약속이 갑자기 잡혔다.	[그녀/NP+와/JKB+의/JKG] (‘와의’ 미등재)

② 조사 결합형이 <우리말샘>에 등재되어 있으면, 사전의 뜻풀이를 참고하여 결합형 자체에 ‘격 조사’나 ‘보조사’라고 풀이되어 있으면 더 분석하지 않고 하나의 조사로 둔다.

[예시] 에다가	[에다가/JKB]
(사전: 일정한 위치를 나타내는 격 조사. 격 조사 ‘에’에 보조사 ‘다가’가 결합한 말이다.)	

③ 만약 조사 결합형이 <우리말샘>에 등재되어 있는데 ‘어떤 조사와 어떤 조사가 결합한 말’로만 풀이되어 있으면 두 개의 조사로 분리하여 분석한다.

[예시] 에는		[에/JKB+는/JX]
	(사전: 부사격 조사 '에'에 보조사 '는'이 결합한 말. 강조와 대조의 뜻을 나타내는 조사이다.)	

1) 격조사(JK)

이는 체언과 다른 성분 간의 일정한 문법 관계를 나타내는 조사이다.

가) 주격조사(JKS)

선행 체언으로 하여금 주어가 되게 하는 조사이다.

이/가	산이 보인다.	[산/NNG+이/JKS]
	우리 둘이 갈게.	[둘/NR+이/JKS]
께서	선생님께서 오신다.	[선생/NNG+님/XSN+께서/JKS]
(이)서	둘이서 그 일을 꾸몄다고?	[둘/NR+이서/JKS]
	혼자서 그 일을 꾸몄다고?	[혼자/NNG+서/JKS]
께오서	부대장님께오서	[부대장/NNG+님/XSN+께오서/JKS]
께옵서	황제께옵서 드나드신다.	[황제/NNG+께옵서/JKS]

주의사항

'이서'의 경우, <우리말샘>에서는 '이'를 접미사로, '서'를 주격조사로 보고 있으나 여기에서는 '이서' 전체를 주격조사로 본다.

주격조사 '이/가'에 대하여 <우리말샘>에서는 '앞말을 지정하여 강조하는 뜻을 나타내는 보조사' 용법을 설정하고 있다. 여기에서는 보조적 연결어미 '-지' 뒤에 나온 '가'만을 보조사로 구별하여 분석한다.

[예시] 예쁘지가 않다.		[예쁘/VA+지/EC+가/JX]

나) 보격조사(JKC)

선행 체언으로 하여금 서술어 '되다, 아니다'의 보어가 되게 하는 조사이다. '되다, 아니다'

앞, 주어가 아닌 요소에 결합한 '이/가'를 보격조사로 분석해야 함에 유의한다.

이/가	얼음이 물이 되었다.	[물/NNG+이/JKC]
	철수는 범인이 아니다.	[범인/NNG+이/JKC]

다) 목적격조사(JKO)

선행 체언으로 하여금 목적어가 되게 하는 조사이다.

ㄹ/을/를	너는 바람소리를 들었다.	[바람/NNG+소리/NNG+를/JKO]

주의사항

목적격조사 'ㄹ/을/를'에 대하여 <우리말샘>에서는 '강조하는 뜻을 나타내는 보조사' 용법을 설정하고 있다. 여기에서는 보조적 연결어미 '-지' 뒤에 나온 'ㄹ/를'만을 보조사로 구별하여 분석한다.

[예시] 밥을 먹질 않는다.	[먹/VV+지/EC+ㄹ/JX]

라) 관형격조사(JKG)

선행 체언으로 하여금 관형어가 되게 하는 조사이다.

의	나의 친구는 너 하나뿐이다	[나/NP+의/JKG]

주의사항

구어에서 '의'가 '에'로 발음되어 '에'로 전사한 경우, 그 '에'는 관형격조사(JKG)로 분석한다.

[예시] 우리에 문제가 바로 그거야.	[우리/NP+에/JKG]

마) 부사격조사(JKB)

선행 체언으로 하여금 부사어가 되게 하는 조사이다.

(으)로	망치로 못을 박아야지.	[망치/NNG+로/JKB]
(으)로서	장관으로서 책임을 다해야 한다.	[장관/NNG+으로서/JKB]
(으)로써	돌로써 지붕을 만든다고?	[돌/NNG+로써/JKB]
같이	바보같이 웃고 다닌다.	[바보/NNG+같이/JKB]
더러	나더러 이것도 하라고 한다.	[나/NP+더러/JKB]
랑	너랑 많이 닮았다.	[너/NP+랑/JKB]
(으)로부터	TV로부터 받는 영향력이 너무 크다.	[TV/SL+로부터/JKB]
마냥	기영이마냥 놀 수만은 없다.	[기영이/NNP+마냥/JKB]
마따나	네 말마따나 나도 그래야 한다.	[말/NNG+마따나/JKB]
만큼	눈물만큼 콧물도 흐른다니까.	[눈물/NNG+만큼/JKB]
보고	영자보고 놀자고 좀 해라.	[영자/NNP+보고/JKB]
보다	직관보다는 논리가 동원돼야 한다.	[직관/NNG+보다/JKB+는/JX]
에	나는 너에 대해 아무것도 모른다.	[너/NP+에/JKB]
에게	너에게 말하기 싫다.	[너/NP+에게/JKB]
에게서	나는 철수에게서 그 말을 들었다.	[철수/NNP+에게서/JKB]
에서	집에서 학교까지 너무 멀다.	[집/NNG+에서/JKB]
에서부터	연구소에서부터 가게까지는 너무 멀다.	[연구소/NNG+에서부터/JKB]
와/과	경미와 함께 다닌다면,	[경미/NNP+와/JKB]
처럼	사람처럼 행동하는 동물이 있다.	[사람/NNG+처럼/JKB]
하고	그 일하고 관련된 사람은 아무도 없다.	[일/NNG+하고/JKB]

바) 호격조사(JKV)

주로 사람을 가리키는 체언 뒤에 연결되어 그것으로 하여금 부름의 대상이 되게 하는 조사이다.

아	호동아! 이제 그만 일어나거라.	[호동/NNP+아/JKV+!/SF]
야	철수야! 밥 먹어라.	[철수/NNP+야/JKV+!/SF]
여	주여, 우리에게 힘을 주소서.	[주/NNG+여/JKV+,/SP]
(이)시여	신이시여! 우리를 저버리지 마소서.	[신/NNG+이시여/JKV+!/SF]

주의사항

호격조사와 어말어미는 구분해서 분석해야 한다.

[예시] 저기 오는 것이 철수야.	[철수/NNP+이/VCP+야/EF+./SF]

사) 인용격조사(JKQ)

인용문이나 인용구를, 동사에 대한 부사적 성분으로 도입하는 조사이다.

고	그는 "이제 가도 좋다"고 말했다.	[좋/VA+다/EF+"/SS+고/JKQ]
(이)라고	"문제가 심각하다"라고 보고했다.	[심각하/VA+다/EF+"/SS+라고/JKQ]

주의사항

① 인용격조사는 연결어미와 구별하기 어려운 경우가 있으므로 주의한다.

[예시] 철수는 자기가 학생이라고 말했다.	[학생/NNG+이라고/JKQ] (×)
	[학생/NNG+이/VCP+라고/EC] (○)
철수는 “다음 주에 놀러 가도 좋다”고 말하였다.	[좋/VA+다/EF+”/SS+고/JKQ] (○)
	[좋/VA+다/EF+”/SS+고/EC] (×)

② 인용격조사는 형태만으로 확인할 수 없고 발화 상황까지 고려해야 하는 복잡한 표지이다. 게다가 인용격조사로 인정되는 형태인 ‘라고’ 등은 원래 용언의 활용형에 불과하다. 하지만 인용격조사를 설정하지 않을 경우에는 인용부호가 들어간 어절의 처리가 어색해진다. 따라서 우리는 인용격조사를 설정하되, 그 쓰임은 인용부호(", ',), },], 〉, ...)가 있는 경우로만 제한하기로 한다. 물론 인용부호가 빠진 경우에는 어미로 분석한다.

[예시] 철수는 영희가 좋다고 말했다.	[좋/VA+다고/EC]

③ 명사 뒤에 따옴표와 '라고/이라고'가 이어지는 경우에도 따옴표 뒤의 '라고/이라고'를 인용격 조사로 분석한다.

[예시] "그것이 우리의 목표"라고 말했다.	[목표/NNG+"/SS+라고/JKQ]

④ 단, 종결어미 + '라고'(직접 인용)의 경우에는 인용부호가 드러나지 않아도 조사로 분석한다.

[예시] 집에 간다라고 했다.	[가/VV+ㄴ다/EF+라고/JKQ] (O)
집에 간다라고 했다.	[가/VV+ㄴ다라고/EC] (×)
[참고] 집에 간다라는 말	[가/VV+ㄴ다/EF+이/VCP+라는/ETM]

⑤ 다음의 경우는 '이/VCP'가 생략된 것이므로 '이/VCP'를 복원하여 분석한다.

[예시] "집에 간다"라는 말에 놀랐다.	[가/VV+ㄴ다/EF+"/SS+이/VCP+라는/ETM]
"바보"라며	["/SS+바보/NNG+"/SS+이/VCP+라며/EC]
"밥을 먹고"라면서	[먹/VV+고/EC+"/SS+이/VCP+라면서/EC]

2) 접속조사(JC)

두 단어를 같은 자격으로 이어 주는 구실을 하는 조사를 말한다. 아래는 그 예시이다.

와	그 아주머니는 딸기와 사과를 샀다.	[딸기/NNG+와/JC]
과	그 기계는 사람과 컴퓨터를 구별하지 못한다.	[사람/NNG+과/JC]
나	사과나 배는 모두 몸에 좋은 과일이다 .	[사과/NNG+나/JC]
랑	머루랑 다래랑 먹으며 청산에 살고 싶어라.	[머루/NNG+랑/JC]
하고	이번 준비물로 칼하고 연필을 샀다.	[칼/NNG+하고/JC]

주의사항

'함께 함'의 뜻을 나타내는 접속조사는 부사격조사와 형태상 동일하므로 주의할 필요가 있다. 체언과 체언 사이에서 두 체언을 이어주는 요소는 접속조사이고, 그 외의 경우에는 부사격조사이다.

[예시]	철수와 영희가 왔다.	[철수/NNP+와/JC]
	철수와 같이 놀았다.	[철수/NNP+와/JKB]

3) 보조사(JX)

체언이나 부사 또는 용언의 연결 어미나 종결 어미의 뒤에 쓰여 특별한 뜻을 더해 주는 조사를 말한다. 아래는 그 예시이다.

그려	먹습니다그려.	[먹/VV+습니다/EF+그려/JX+./SF]
까지(꺼정/까장)	너까지 나에게 이럴 줄이야.	[너/NP+까지/JX]
깨나	너도 사람깨나 울렸겠구나.	[사람/NNG+깨나/JX]
(이)나	너나 가라!	[너/NP+나/JX]
(이)나마	빵이나마 먹어라.	[빵/NNG+이나마/JX]
ㄴ/은/는	이 종이는 어제 사 온 것이다.	[종이/NNG+는/JX]
ㄴ커녕/은커녕/는커녕	돈은커녕 먹을 쌀도 없다.	[돈/NNG+은커녕/JX]
다	그 물건을 거기다 놓아라.	[거기/NP+다/JX]
다가	책상을 어디다가 둘까요?	[어디/NP+다가/JX]
대로(대루)	너는 너대로 살아라.	[너/NP+대로/JX]
따라	오늘따라 택시도 안 잡힌다.	[오늘/NNG+따라/JX]
도/두	강아지도 주인은 알아본다.	[강아지/NNG+도/JX]
(이)란	코알라란 호주에 사는 초식동물이다.	[코알라/NNG+란/JX]
만	인간은 빵만으로 살 수 없다.	[빵/NNG+만/JX+으로/JKB]
밖에	그래 봐야 죽기밖에 더 하랴.	[죽/VV+기/ETN+밖에/JX]
부터/부텀	우선 노인부터 태워라.	[노인/NNG+부터/JX]
뿐	가진 건 고작 집 한 채뿐.	[채/NNB+뿐/JX]
(이)야	그가 인간성이야 그만이지.	[인간/NNG+성/XSN+이야/JX]
요	나는요 그림을요 예쁘게 그립니다.	[나/NP+는/JX+요/JX]

조차	이젠 집조차 빼앗기는구나.	[집/NNG+조차/JX]
치고	값싼 물건치고 쓸 만하다.	[물건/NNG+치고/JX]

(1) 보조사 분석 기준

앞에 '이'가 개재될 수 있는 조사는 지정사 '이다'에 어미가 결합한 형태와 구분하기 어려운 경우가 있다. 본 분석에서는 <우리말샘>을 따라 '이'형 조사와 지정사 '이다'의 활용형을 구분하는 것을 원칙으로 한다. '라든지'처럼 <우리말샘>에서 조사로만 다루어지는 것은 조사로 처리하면 되지만, '든지'처럼 <우리말샘>에서 조사와 어미 모두로 등재되어 있는 것은 문맥과 <우리말샘>의 예문을 참조하여 조사인지 '이다'의 활용형인지를 판단해야 한다. 조사와 '이다' 활용형의 기본적인 구별 기준은 다음과 같다.

(가) '이-' 뒤에 '-시-'나 '-었-' 등의 선어말어미가 결합할 수 있으면 그 뒤의 요소는 어미이다. '이-' 뒤에 선어말어미가 결합할 수 없으면 전체가 '이'를 포함하는 조사이다.

[예시] 학생이라도 지원할 수 있습니다. [학생/NNG+이/VCP+라도/EC]
cf) 학생이시라도 지원하실 수 있습니다.
cf) [철수가 학생이라도] 지원할 수 있습니다.
[예시] 사람이 부족하니 선생님이라도 빨리 오세요. [선생/NNG+님/XSN+이라도/JX]
cf) *선생님이시라도 빨리 오세요.
cf) *[당신이 선생님이라도] 빨리 오세요.

(나) '체언+이-'의 주어를 상정할 수 있으면 그 뒤의 요소는 어미이다. '체언+이-'의 주어를 상정할 수 없으면 전체가 '이'를 포함하는 조사이다.

(다) 다음의 형태는 지정사 '이다'의 활용형과는 관계가 없으므로 모두 보조사가 된다.

[예시] 까지, 깨나, 는(은/ㄴ), 대로, 도, 따라, 마다, 마저, 만, 밖에, 부터, 뿐, 조차, 치고, ㄴ커녕

주의사항

① 다음의 형태들은 분석 결과에 중의성이 생기므로, 이들을 분석할 때는 특히 주의해야 한다.

[예시]	(이)란 코알라란 동물은 호주에 주로 서식한다.	[코알라/NNG+이/VCP+란/ETM]
	코알라란 매우 귀여운 동물이다.	[코알라/NNG+란/JX]
	(이)나 밥이나 빵을 먹도록 해라.	[밥/NNG+이나/JC]
	밥이나 먹자.	[밥/NNG+이나/JX]
	그가 비록 열심히 하나 능력은 부족하다.	[하/VV+나/EC]
	어제 내가 술을 마셨나?	[마시/VV+었/EP+나/EF+?/SF]
	(이)야 철수야 당연히 그 일을 할 수 있지.	[철수/NNP+야/JX]
	내가 좋아하는 것은 철수야.	[철수/NNP+이/VCP+야/EF+./SF]
	철수야! 부르는 소리	[철수/NNP+야/JKV+!/SF]
	(이)요 밥을 먹다가요	[먹/VV+다가/EC+요/JX]
	밥이요 빵이요	[밥/NNG+이/VCP+요/EC]

② 구어에서 받침 있는 말 뒤에서 '요' 대신 쓰이는 '이요'는 보조사로 분석한다.

[예시]	A: 넌 머 먹을래? B: 전 밥이요.	[밥/NNG+이요/JX+./SF]

③ '종결어미+요(보조사)'는 <우리말샘>에 등재되어 있는 '어요, 지요, 래요' 등을 제외하고 모두 원래의 범주인 종결어미와 보조사로 분리하여 분석한다.

[예시]	우리 집에 갈까요?	[가/VV+ㄹ까/EF+요/JX+?/SF]
	어디서 저녁 먹나요?	[먹/VV+나/EF+요/JX+?/SF]
	빨리 공부해야지요.	[공부/NNG+하/XSV+아야지/EF+요/JX+./SF]

④ '비종결어미+요(보조사)'는 통합하지 않고 각각 분석해 준다.

[예시]	제가 몸이 좀 아파서요 지각을 했어요.	[아프/VA+아서/EC+요/JX]
	내가요, 왜요	[내/NP+가/JKS+요/JX], [왜/MAG+요/JX]

⑤ '말고'는 용언 '말다'의 활용형으로 처리한다.

[예시]	돈말고 지혜가 필요하다.	[돈/NNG+말/VV+고/EC]

바 의존형태

1) 어미(E)

가) 선어말어미(EP)

용언이 활용할 때, 어간과 어말 어미 사이에 나타나는 것으로 높임법이나 시제, 양태를 나타내는 문법적인 요소이다. 선어말어미의 목록은 연구자에 따라 다를 수 있으나 이 분석에서는 아래의 것만을 선어말어미로 인정한다.

-겠-	그 일은 내일 처리하겠다.	[처리/NNG+하/XSV+겠/EP+다/EF+./SF]
-(으)시-	선생님께서 손수 만드신	[만들/VV+시/EP+ㄴ/ETM]
-옵-	어머님께 선물을 바치옵고	[바치/VV+옵/EP+고/EC]
-았/었-	우리가 먹었던 음식에 문제가 있다.	[먹/VV+었/EP+던/ETM]
-았었/었었-	거기는 우리가 전에 갔었던 곳이야.	[가/VV+았었/EP+던/ETM]

주의사항

① 어간 '하-' 뒤에 과거 시제 선어말어미가 결합하여 '했'의 형태로 나타나거나 '하였'의 형태로 나타날 수 있는데, 본 분석에서는 이 경우 '-였-' 형태를 인정하지 않고 모두 '-았-'으로 분석한다. 이 외에 '아/어'를 포함하고 있는 모든 어미 역시, '하-' 뒤에 나타난 경우에는 '아X'형으로 분석한다.

② 다음의 선어말어미는 그 어간이 생략되었을 경우에 어간을 복원해 준다.

-겠-	이것은 그대로 두어야겠다.	[두/VV+어야/EC+하/VX+겠/EP+다/EF+./SF]
-았/었-	철수가 그것을 가져오랬다.	[가져오/VV+라/EF+하/VV+았/EP+다/EF+./SF]
-시-	선생님께서 가자시오.	[가/VV+자/EF+하/VV+시/EP+오/EF+./SF]

③ 위의 선어말어미가 포함되지 않은 어미 형태는 그대로 어미로 분석한다.

-랄까, -대야, -래야

나) 종결어미(EF)

용언의 어간이나 선어말어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 한 문장을 끝맺는 역할을 한다. 본 지침에서는 <우리말샘>에 따라 종결어미를 구분한다. 다음은 종결어미의 일부 사례이다.

-게	그만한 돈이 있으면 좋게.	[좋/VA+게/EF+./SF]
-ㄴ가	이것이 무엇인가?	[무엇/NP+이/VCP+ㄴ가/EF+?/SF]
-ㄴ걸	이제 시작인걸.	[시작/NNG+이/VCP+ㄴ걸/EF+./SF]
-ㄴ다	이건 말도 안 된다.	[되/VV+ㄴ다/EF+./SF]
-나	자네 그리로 가나?	[가/VV+나/EF+?/SF]
-냐	키가 얼마나 크냐?	[크/VA+냐/EF+?/SF]
-네	정말 큰일 났네!	[나/VV+았/EP+네/EF+!/SF]
-는걸	그는 벌써 갔는걸.	[가/VV+았/EP+는걸/EF+./SF]
-는구나	앞이 잘 안 보이는구나.	[보이/VV+는구나/EF+./SF]
-는구려	잘도 먹는구려.	[먹/VV+는구려/EF+./SF]
-는구먼	공부를 잘하는구먼.	[잘/MAG+하/XSV+는구먼/EF+./SF]
-는다	아이가 글을 잘 읽는다.	[읽/VV+는다/EF+./SF]
-다	그게 사실이다.	[사실/NNG+이/VCP+다/EF+./SF]
-ㄹ게	그렇게 할게.	[하/VV+ㄹ게/EF+./SF]
-ㅂ니까	이제야 옵니까?	[오/VV+ㅂ니까/EF+?/SF]
-ㅂ니다	이렇게 합니다.	[하/VV+ㅂ니다/EF+./SF]
-습니까	그래도 되겠습니까?	[되/VV+겠/EP+습니까/EF+?/SF]
-습니다	정말 재미있습니다.	[재미있/VA+습니다/EF+./SF]
-ㅂ시다	다시 만납시다.	[만나/VV+ㅂ시다/EF+./SF]
-ㅂ시오	서둘러 주십시오.	[주/VX+시/EP+ㅂ시오/EF+./SF]
-으냐	물이 얼마나 깊으냐?	[깊/VA+으냐/EF+?/SF]
-은가	그것이 좋은가?	[좋/VA+은가/EF+?/SF]

-오/으오/소	물이 깨끗하오.	[깨끗하/VA+오/EF+./SF]
-ㅂ디다/습디다	참 좋은 곳입디다.	[곳/NNB+이/VCP+ㅂ디다/EF+./SF]
-거든	나는 이것이 좋거든!	[좋/VA+거든/EF+!/SF]
-ㄴ걸/은걸	힘이 꽤 센걸.	[세/VA+ㄴ걸/EF+./SF]
-ㄹ걸/을걸	모른다고 할걸.	[하/VV+ㄹ걸/EF+./SF]
-ㄹ까	이제 밥을 할까?	[하/VV+ㄹ까/EF+?/SF]
-다오	그가 가지고 있다오.	[있/VX+다오/EF+./SF]
-다네	일을 망쳤다네	[망치/VV+었/EP+다네/EF+./SF]
-다구	돈이 많다구?	[많/VA+다구/EF+?/SF]
-다니까	돈이 없다니까!	[없/VA+다니까/EF+!/SF]
-냐고/느냐고	그가 누구냐고?	[누구/NP+이/VCP+냐고/EF+?/SF]
-도다	꽃이 아름답도다.	[아름답/VA+도다/EF+./SF]
-다니	그가 책을 읽다니!	[읽/VV+다니/EF+!/SF]
-는가	같이 가겠는가?	[가/VV+겠/EP+는가/EF+?/SF]
-ㅂ디까/습디까	보기에 좋습디까?	[좋/VA+습디까/EF+?/SF]
-다면서	술은 싫다면서?	[싫/VA+다면서/EF+?/SF]
-다나	그도 가겠다나.	[가/VV+겠/EP+다나/EF+./SF]
-렴/으렴	맘대로 해 보렴.	[보/VX+렴/EF+./SF]
-려무나	책이나 읽으려무나.	[읽/VV+으려무나/EF+./SF]
-라니까	그 사람이 아니라니까.	[아니/VCN+라니까/EF+./SF]
-세	일이나 하세.	[하/VV+세/EF+./SF]
-자꾸나	약속을 좀 미루자꾸나.	[미루/VV+자꾸나/EF+./SF]
-자니까	그만 따지자니까.	[따지/VV+자니까/EF+./SF]
-아/어/야	밥 먹어!	[먹/VV+어/EF+!/SF]
-ㅁ세/음세	그날 꼭 옴세.	[오/VV+ㅁ세/EF+./SF]
-단다	애들이 다쳤단다.	[다치/VV+었/EP+단다/EF+./SF]
-더라고	아까 보니 철수가 집에 가더라고.	[가/VV+더라고/EF+./SF]

주의사항

① '종결어미+요(보조사)'는 <우리말샘>에 등재되어 있는 '어요, 지요, 래요' 등을 제외하고 모두 종결어미와 보조사로 분리하여 분석한다.

[예시] 말씀대로 했는걸요.	[하/VV+았/EP+는걸/EF+요/JX+./SF]

② '-세요'는 다음과 같이 선어말어미까지 분석한다.

[예시] 어서 출근하세요.	[출근/NNG+하/XSV+시/EP+어요/EF+./SF]

③ '-죠'는 축약형을 그대로 태깅한다. 단, 종결어미 '-어야지'와 '요'가 결합하여 '-어야죠' 형식으로 나왔을 때는 '-지'와 '요'를 분리한다.

[예시] 어서 출근하죠.	[출근/NNG+하/XSV+죠/EF+./SF]
어서 출근해야죠.	[출근/NNG+하/XSV+아야지/EF+요/JX+./SF]

④ "앞의 사실을 청자가 이미 알고 있음"을 나타내는 '잖'은, '-더-', '-는-' 등과 같이 본 지침상 분석하지 않는 선어말어미처럼 취급하여 다음과 같이 어말어미와 결합하여 표지를 부여한다.

[예시] 저 오늘 일찍 일어났잖아요.	[일어나/VV+았/EP+잖아요/EF+./SF]
제가 말씀드렸잖습니까.	[말씀드리/VV+었/EP+잖습니까/EF+./SF]

단, '하' 생략과 함께 '지 않'이 줄어들어서 나타난 '잖' 형은 '하'와 함께 '지 않'을 복원하여 분석한다. 이때의 '잖'은 용언 어간이나 선어말어미 뒤에 붙어 "앞의 사실을 청자가 이미 알고 있음"을 나타내는, 선어말어미에 준하는 '잖'이 아님에 유의한다.

[예시] 녹록잖은 일이다.	[녹록하/VA+지/EC+않/VX+은/ETM]

⑤ '-려고'는 <우리말샘>에서 의심과 반문의 용법으로만 종결어미 자격을 갖는 것으로 등재되어 있다. 하지만 아래와 같이 뒤에 생략된 말 없이 주어의 의도만을 밝히며 문말에서 쓰이는 '-려고'는 종결어미 용법으로 볼 수 있으므로 종결어미로 분석한다.

[예시] 나는 오늘 집에 일찍 가려고. [가/VV+려고/EF+./SF]

→ '-려고' 뒤에 '생각하다' 정도의 동사가 생략되어 있다. 이때는 주어의 의도가 무엇인지만을 밝히며 문말에서 쓰인 '-려고'로 볼 수 있으며, 종결어미로 분석한다.

[예시] 나 요즘 매일 운동해. 살 빼려고. [빼/VV+려고/EC+./SF]

→ '-려고' 뒤에 '생각하다'가 생략된 것이 아니며 '운동하다'와 같이 주어의 의도를 실현하기 위한 행동을 나타내는 말이 생략되어 있다. 이때는 '-려고'가 연결어미로 쓰인 것이다.

다) 연결어미(EC)

용언의 어간이나 선어말어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 문장을 종결시키지 못하고 뒤에 오는 절을 연결시켜 주는 어미를 말한다. 본 지침에서는 <우리말샘>에 따라 연결어미를 구분하는 것을 원칙으로 한다. 다음은 연결어미의 일부 사례이다.

-거나	누가 오거나 알은 체 할 것 없다.	[오/VV+거나/EC]
-거든	거기 가거든 김 사장이 있는지 보아라.	[가/VV+거든/EC]
-건대	내가 보건대, 네 말이 옳다.	[보/VV+건대/EC]
-건마는	말렸건마는 아직도 축축하다.	[말리/VV+었/EP+건마는/EC]
-게	개를 굶게 하지 마라.	[굶/VV+게/EC]
-고	일을 하고 밥을 먹자.	[하/VV+고/EC]
-곤	숙제한 것도 빌려가곤 한다.	[빌리/VV+어/EC+가/VV+곤/EC]
-기에	늦게라도 왔기에 용서해 주었다.	[오/VV+았/EP+기에/EC]
-ㄴ다기에	잠시 쉰다기에 승낙했다.	[쉬/VV+ㄴ다기에/EC]
-ㄴ다손/다손	밉다손 치더라도 구박하지 말자.	[밉/VA+다손/EC]
-ㄴ들/는들	간다 한들 아주 갈까?	[하/VV+ㄴ들/EC]
-ㄴ즉	배가 고픈즉 속이 쓰리다.	[고프/VA+ㄴ즉/EC]
-ㄴ지라/는지라	눈이 온지라 길이 미끄럽다.	[오/VV+ㄴ지라/EC]
-나	눈이 오나 비가 오나 같다.	[오/VV+나/EC]
-나마	맛이 좋지 못하나마 많이 드십시오.	[못하/VX+나마/EC]
-는다기에	빵을 먹는다기에 주었다.	[먹/VV+는다기에/EC]
-니까	너를 보니까 좋다.	[보/VV+니까/EC]

-다가	자랑하다가 망신당했다.	[자랑/NNG+하/XSV+다가/EC]
-다기에	꽃이 예쁘다기에 보러 왔소.	[예쁘/VA+다기에/EC]
-대도	시간이 있대도 만나 주질 않는다.	[있/VA+대도/EC]
-더라도	가더라도 꼭 돌아와라.	[가/VV+더라도/EC]
-던들	진작 알았던들 방법을 취했지.	[알/VV+았/EP+던들/EC]
-든지	외모가 어떠하든지 무슨 상관인가?	[어떠하/VA+든지/EC]
-ㄹ뿐더러	비가 올뿐더러 바람도 분다.	[오/VV+ㄹ뿐더러/EC]
-ㄹ수록	높이 올라갈수록 춥다.	[올라가/VV+ㄹ수록/EC]
-ㄹ지	비가 얼마나 올지 천둥이 다 친다.	[오/VV+ㄹ지/EC]
-ㄹ지라도	이길지라도 명예롭지는 않다.	[이기/VV+ㄹ지라도/EC]
-ㄹ지언정	죽을지언정 그 일은 못하겠다.	[죽/VV+을지언정/EC]
-라고	철수는 자기가 바보라고 생각한다.	[바보/NNG+이/VCP+라고/EC]
-락	자락 깨락 잠을 설쳤다.	[자/VV+락/EC]
-랍시고	그는 반장이랍시고 거드름만 피운다.	[반장/NNG+이/VCP+랍시고/EC]
-려니와	비용도 문제려니와 일꾼도 문제다.	[문제/NNG+이/VCP+려니와/EC]
-련마는	보면 반가우련마는 볼 수가 없네.	[반갑/VA+으련마는/EC]
-면	지옥이 존재하면 만원일 것이다.	[존재/NNG+하/XSV+면/EC]
-면서	푸르면서 검은 물빛	[푸르/VA+면서/EC]
-므로	비가 오므로 가지 않겠다.	[오/VV+므로/EC]
-아/어	입을 막아 버렸다.	[막/VV+아/EC]
-아도/어도	암만 봐도 모르겠다.	[보/VV+아도/EC]
-아서/어서	덫을 놓아서 꿩을 잡았다.	[놓/VV+아서/EC]
-아야	이 일은 잘해야 한다.	[잘/MAG+하/XSV+아야/EC]
-으나	밥을 먹으나 마나이다.	[먹/VV+으나/EC]
-으나마	맛은 없으나마 많이 드세요.	[없/VA+으나마/EC]
-자마자	집에 오자마자 씻었다.	[오/VV+자마자/EC]
-지	밥을 먹지 못했다.	[먹/VV+지/EC]
-지마는	비가 오지마는 가야 한다.	[오/VV+지마는/EC]

주의사항

① 어미에 따라서는 분석의 중의성이 생길 수 있으므로 문맥 확인을 통해 형태분석을 결정한

다.

[예시]	너는 내가 왔는데 기쁘지도 않니?	[오/VV+았/EP+는데/EC]
	철수가 있는데가 어디지?	[있/VA+는/ETM+데/NNB+가/JKS]
	다들 만족하는지 아무런 불평이 없다.	[만족/NNG+하/XSV+는지/EC]
	다들 만족하는지는 모르겠다.	[만족/NNG+하/XSV+는지/EF+는/JX]
	너를 만난지도 꽤 오래구나.	[만나/VV+ㄴ/ETM+지/NNB+도/JX]

'-을까', '-는가, -은가'는 언제나 종결어미임에 유의한다. '누군가', '어딘가' 등도 '누구/NP+이/VCP+ㄴ가/EF', '어디/NP+이/VCP+ㄴ가/EF' 등으로 분석한다.

'-을지', '-는지, -은지'는 연결어미 용법과 종결어미 용법을 모두 갖는데, 뒤에 조사가 오거나 '모르다'의 목적어 자리에서 쓰이는 경우 종결어미 용법에 해당한다는 점에 유의한다.

② 통사적 구성에 나타나는 '-음직'은 '음직/EC'로 분석한다. 그러나 '바람직하다, 먹음직하다' 등 사전에 등재되어 있는 단어의 내부에서 확인되는 '-음직'은 더 이상 분석할 수 없다는 것에 유의한다.

[예시]	철수라면 외국에 갔음직 하다.	[가/VV+았/EP+음직/EC 하/VA+다/EF+./SF]
	어른답고 믿음직하게 행동해라.	[믿음직하/VA+게/EC]
	그것 참 먹음직스럽다.	[먹음직스럽/VA+다/EF+./SF]
	그것은 매우 바람직한 일이다.	[바람직하/VA+ㄴ/ETM]

라) 명사형전성어미(ETN)

한 문장의 성격을 임시로 바꾸어 다른 문장 속에서 명사적인 역할을 하게 하는 어미를 말한다.

-기	그 일은 정말 중요하기 때문이다.	[중요/NNG+하/XSA+기/ETN]
-ㅁ/-음	장사는 신용을 얻음이 제일이다.	[얻/VV+음/ETN+이/JKS]

주의사항

① ㅂ 불규칙 용언 어간에 명사형 전성 어미 '-음'이 결합한 경우 '-음'이 아닌 '-ㅁ'으로 분석한다.

ㅅ 불규칙 용언 어간에 결합하는 '-음'은 '음/ETN'으로 분석한다.

[예시] 아니꼬움을 견디지 못하고	[아니꼽/VA+ㅁ/ETN]
[예시] 김철수 지음	[짓/VV+음/ETN]

② '음, 기'가 붙은 말이 단순히 명사형이냐 아니면 굳어진 명사이냐 하는 것은 물론 문맥에 따라 결정되어야 하지만 먼저 그것이 사전에 등재되어 있느냐의 여부를 살펴보아야 한다.

[예시] 책을 읽기가 어렵다.	[읽/VV+기/ETN+가/JKS]
읽기 교육이 문제가 된다.	[읽기/NNG]

③ 여러 개의 어미가 결합한 준말의 끝에 명사형 전성 어미가 나오는 다음과 같은 경우, 어미를 모두 묶어서 명사형 전성 어미로 표지를 부여한다.

[예시] 꼭 그렇다기보다는	[그렇/VA+다기/ETN+보다/JKB+는/JX]
그것이 문제라기에는	[문제/NNG+이/VCP+라기/ETN+에/JKB+는/JX]

마) 관형형전성어미(ETM)

용언의 성격을 임시로 바꾸어 다른 문장 속에서 관형사적인 역할을 하게 하는 어미이다.

-ㄴ/은	어제 먹은 빵에 이상이 있었다.	[먹/VV+은/ETM]
-는	잃어버린 물건을 찾는 일은 어렵다.	[찾/VV+는/ETM]
-던	이제까지 미루던 일을 오늘 해치웠다.	[미루/VV+던/ETM]
-ㄹ/을	나에게는 아직 처리할 일이 있다.	[처리/NNG+하/XSV+ㄹ/ETM]
-런	우리가 함께한 날이 어제런 듯하다.	[어제/NNG+이/VCP+런/ETM]

주의사항

① ㅂ 불규칙 용언 어간에 관형사형 전성 어미가 결합한 경우 '-은, -을'이 아닌 '-ㄴ, -ㄹ'로 분석한다. 이는 '-ㄴ, -ㄹ'을 포함하고 있는 '-ㄴ가', '-ㄹ까' 등에도 적용된다. 이러한 방식은 모든 불규칙 용언과 모든 매개모음 어미에 적용되는 것이 아니라, ㅂ 불규칙 용언 어간에 명사형 어미(-ㅁ)와 관형사형 어미(-ㄴ, -ㄹ)가 결합할 때 적용됨에 유의한다. ㅅ 불규칙 용언 어간에 결합하는 '-은, -을'은 '은/ETM, 을/ETM'으로 분석한다.

[예시] 그녀의 고운 얼굴	[곱/VA+ㄴ/ETM]
꽃밭은 매우 아름다울 것이다.	[아름답/VA+ㄹ/ETM]
얼마나 고울까?	[곱/VA+ㄹ까/EF+?/SF]
[참고] 얼굴이 고우니	[곱/VA+으니/EC]
[참고] 집을 지을 거야.	[짓/VV+을/ETM]

② 종결어미에 이어서 전성어미가 올 경우 통합해서 전성어미로 처리한다.

[예시] 어느 쪽에 더 비중을 두느냐는 것이	[두/VV+느냐는/ETM]

2) 접사(X)

주의사항

접사(접두사, 접미사)는 아래 가)~라)에 목록화된 접사가 등장한 경우에만 분리하여 분석한다. 접사의 분리 원칙은 다음과 같다.

2음절 단어의 처리

① <우리말샘> 등재어인 경우(즉 원어절에 나타난 2음절 단어와 같은 의미의 단어가 <우리말샘> 표제어로 올라 있는 경우), 표제어에 하이픈이 있는 경우에만 접사를 분리한다. 단, 접사를 분리하고 남는 요소가 어근에 해당하는 경우에는 접사를 분리하지 않는다.

[예시] 오형 (사전: 오-형, 혈액형의 하나)	[오/NNG+형/XSN]

<우리말샘>에 명사, 부사로 등재된 의미나 쓰임이 아니라, 그 앞뒤로 다른 말(단위성 의존명사 등)과 함께 쓰이는 '순서'의 '제일'은 '제/XPN+일/NR'로 분리한다.

[예시] 제일 차(회/조/항…) 회의	[제/XPN+일/NR] (O) [제일/NNG] (×), [제일/MAG] (×)

분석 대상이 되는 말이 <우리말샘>에 등재되어 있다고 하더라도 그 쓰임과 의미를 면밀히 확인할 수 있도록 주의한다.

[예시] 15일자 신문	[15/SN+일/NNB+자/NNG] (○) [15/SN+일자/NNG] (×)

② <우리말샘> 미등재어인 경우, 그 단어가 2음절 한자어이거나 접사 분리 시 어근이 남는다면 접사를 분리하지 않는다. 그 외의 경우에는 접사를 분리한다.

[예시] 뇌성마비(사전: 뇌성^마비) [뇌성/NNG+마비/NNG]

→ <우리말샘>에서 '뇌성'은 단독 표제어로 올라 있지 않아 하이픈 유무를 참고할 수 없다. 하지만 2음절 한자어에 해당하는 말이 <우리말샘>에서 대체로 하이픈 없이 처리되고 있음을 참고하여 '-성'을 분리하지 않고 명사로 처리한다.

[예시] 나는 아침형 인간이 아니라 밤형 인간이다. [밤/NNG+형/XSN]

→ '밤형'은 미등재어이지만 2음절 한자어가 아니다. 또한 '밤'이 명사이므로 '밤'과 '-형'을 분리한다.

③ 숫자, 로마자 등 기타기호에 접사가 결합한 것은 일반명사 지침의 (라)항을 우선 적용하여 처리한다.

[예시] 3분의 일 [3/SN+분/XSN]

→ 명사 '삼분'이 <우리말샘>에 하이픈 없이 등재되어 있지만, 기타기호의 처리 방법을 우선 적용하여 [3/SN+분/XSN]으로 분리하여 분석한다.

3음절 이상 단어의 처리

① 3음절 이상 복합어의 경우, <우리말샘>의 표제어 하이픈(-) 위치를 참고하여 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있는 경우에만 해당 접사를 분리한다.

[예시] 과보호(사전: 과-보호) [과/XPN+보호/NNG]

→ 하이픈 바로 앞에 놓인 접사인 '과-'가 분석 대상 접사이다. 이 경우 '과'와 '보호'를 분리한다.

[예시] 피보험자(사전: 피보험-자) [피보험자/NNG]

→ 하이픈 바로 뒤에 놓인 접사인 '-자'는 본 지침의 분석 대상 접사가 아니다. 따라서 더 이상 분리하지 않고 전체를 [피보험자/NNG]로 분석한다.

② 만약 접사를 분리했을 때 남는 단위가 어근(XR)이라면 접사 분리를 하지 않는다.

[예시] 비롯하다(사전: 비롯-하다) [비롯하/VV+다/EF]

→ 하이픈 바로 뒤에 놓인 접사인 '-하-'가 분석 대상 접사이지만, 이것을 분리하고 남는 단위인 '비롯'이 어근에 해당한다. 이 경우 '비롯'과 '하'를 분리하지 않는다.

③ 접사를 분리하고 남은 부분이 사전 미등재어인 경우가 있다. 그 미등재어가 홀로 쓰이지 않아 어근 자격을 갖는 것으로 판단된다면, 위 ②와 마찬가지로 접사를 분리하지 않는다.

[예시] 역세권(사전: 역세-권) [역세권/NNG]

→ 하이픈 바로 뒤에 놓인 접사인 '-권'이 분석 대상 접사이지만, 이것을 분리하고 남는 단위인 '역세'가 사전 미등재어이며 홀로 쓰이지도 않아 어근에 해당한다. 이 경우 '역세'와 '권'을 분리하지 않는다.

접사를 분리하고 남은 미등재어가 합성어라면, 합성어 분석 원칙에 따라 합성어 구성 요소를 분리하여 분석한다.

[예시] 중고생(사전: 중고-생) [중/NNG+고/NNG+생/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-생'이 분석 대상 접사인데, 이것을 분리하고 남는 단위인 '중고'가 사전 미등재어이다. 그런데 사전에 중학교를 뜻하는 '중', 고등학교를 뜻하는 '고'가 명사로 올라 있어 이 말은 합성어로 파악된다. 이 경우 접사를 분리하고 남은 미등재 합성어를, 여타 미등재 합성어의 처리 방식과 마찬가지로 분리하여 분석한다.

[참고] 일회용(사전: 일회-용) [일회/NNG+용/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-용'이 분석 대상 접사인데, 이것을 분리하고 남는 단위인 '일회'가 단독으로는 사전 등재어가 아니다. 하지만 '일회^결실성' 등의 구 표제어 속에서 한 단어로 나타나므로 더 분석하지 않고 한 단어로 취급한다.

④ 만약 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있어서 해당 접사를 분리해 냈는데, 접사를 뗀 나머지 부분에 또 분석 대상 접사가 포함되어 있을 수 있다. 그런 경우에는 그 나머지 단어를 <우리말샘>에서 검색하여 하이픈의 위치를 확인한 후, 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있다면 해당 접사를 다시금 분리해 낸다.

[예시] 비합리적(사전: 비합리-적) [비/XPN+합리/NNG+적/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-적'이 분석 대상 접사이므로 '비합리'와 '적'을 분리한다. 그런데 '적'을 뗀 나머지 부분인 '비합리'(사전: 비-합리)에 분석 대상 접사인 '비'가 들어 있고, <우리말샘>에서 '비합리'를 검색했을 때 하이픈 바로 앞에 '비'가 놓여 있다. 이 경우 '비'와 '합리'를 다시금 분리한다. 결과적으로 '비합리적'을 [비/XPN+합리/NNG+적/XSN]으로 분석하게 된다.

⑤ 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있어서 해당 접사를 분리했을 때 남는 단위가 2음절 요소라면 위 '2음절 단어의 처리'에 따라 해당 2음절 요소를 처리한다.

⑥ 복합어가 <우리말샘> 미등재어여서 하이픈 정보를 참고할 수 없는 경우에는 <우리말샘> 등재 어휘를 참조하고 복합어의 의미 구조에 대해 직접 판단하여 처리한다.

[예시] 최대형 (미등재어) [최대/NNG+형/XSN]

→ 사전 등재어인 '최소형'(사전: 최소-형)을 참고하여 처리할 수 있다.

[예시] 대의원회 (미등재어) [대의원회/NNG]

→ '대의원의 모임'이라는 뜻이므로 의미 구조상 대의원-회로 나뉜다. 이때 하이픈 뒤의 '-회'는 본 지침의 분석 대상 접사가 아니므로 이 단어를 더 분리하지 않는다.

가) 체언접두사(XPN)

접두사는 명사와 수사에 결합하는 접사류를 묶어서 체언접두사만을 설정하기로 한다.

명사 접두사에는 한자어계 접두사와 고유어계 접두사가 있는데, 그 목록의 풍부함에 비해 대개가 생산성이 그리 높지 않다. 일단 여기서는 비교적 생산성이 높다고 인정되는 접두사에 대해 접두사 분석을 하기로 한다.

하나의 접두사가 여러 개의 다의를 갖는 경우가 있다. 아래에 제시한 접두사가 하나의 표제어 안에서 다의를 갖는 경우, 어떤 의미로 쓰인 것이든 관계없이 접두사를 분석해 낸다. ('친(親)-'이 "혈연관계로 맺어진"의 뜻으로 쓰였든 "부계 혈족 관계인"의 뜻으로 쓰였든 "그것에 찬성하는"의 뜻으로 쓰였든 모두 접두사 분석을 한다.)

가(假)	가건물	소(小)	소강당
고(高)	고물가	신(新)	신정당
과(過)	과보호	왕(王)	왕족발
구(舊)	구소련	재(再)	재충전
날	날음식	저(低)	저임금
노(老)	노부부	제(第)	제13차
대(大)	대선배	준(準)	준전시
맏	맏아들	초(超)	초만원
맨	맨몸	최(最)	최고급
무(無)	무의식	친(親)	친러시아
미(未)	미완성	탈(脫)	탈냉전
반(反)	반독재	폐(廢)	폐광산
범(汎)	범세계	풋	풋살구
부(不)	부도덕	피(被)	피고소인
불(不)	불합리	한	한가운데
비(非)	비논리	헛	헛고생
생(生)	생김치		

나) 명사파생접미사(XSN)

명사파생접미사는 명사나 다른 어근에 후행하여 그것이 명사의 기능을 수행할 수 있도록 만들어 주는 의존 형태이다. 그러나 명사파생접미사는 연구자에 따라 그 목록이 다르며, 실제로도 구분이 애매한 경우가 많다. 본 분석에서는 접미사의 생산성과 접미사를 제외한 형태의 독립성을 기준으로 다음과 같이 목록을 마련하였다.

하나의 접미사가 여러 개의 다의를 갖는 경우가 있다. 아래에 제시한 접미사가 하나의 표제어 안에서 다의를 갖는 경우, 어떤 의미로 쓰인 것이든 관계없이 접미사를 분석해 낸다.

가(哥)	김가	분지(分之)	삼분지 일
가(價)	매매가	뻘	조카뻘
가량	1시간가량, 다섯명가량	산(産)	중국산
간(間)	한 달간	상(上)	역사상
경(頃)	두 시경	생1(生)	갑자생
계(界)	교육계	생2(生)	견습생
계(系)	몽고계	성(性)	인간성
광(狂)	메모광	시(視)	영웅시
권(券)	만 원권	씩	만원씩
권(圈)	운동권	어치	만원어치
권(權)	참정권	여(餘)	삼십여
기(氣)	기름기	용(用)	전쟁용
께	10분께	율(率)	출산율
꼴	십 원꼴	장이	간판장이
꾼	노름꾼	쟁이	심술쟁이
끼리	전우끼리	적(的)	사상적
네	동이네	정(整)	일만 원정
님	선생님	제(制)	봉건제
당(當)	한 사람당	질	서방질
대(臺)	만 원대	짜리	백 원짜리
댁(宅)	청주댁	째1	이틀째
들	우리들	째2	옹기째
들이	1ℓ들이	쯤	내일쯤
론(論)	비평론	층(層)	선수층
류(類)	자연류	치(値)	기대치
률(率)	경쟁률	치레	인사치레
리(裡)	비밀리	투성이	먼지투성이
발(發)	서울발	풍(風)	복고풍
배기	열 살배기	하(下)	지배하
별(別)	가구별	형(型)	기본형
부(附)	12일부	형(形)	계란형
분(分)	3분의 일	화(化)	도구화

주의사항

① 명사파생접미사인 '-들'은 그 분포가 매우 다양하여 일부에서는 이를 보조사와 접미사로 나누어 분석하기도 한다. 그러나, 본 분석에서는 이들을 모두 명사파생접미사로 처리한다. '먹고들'의 '-들'도 선행성분이 어미이긴 하나, 일치하는 대상은 선행하는 명사로 해석할 수도 있기 때문이다.

[예시] 사람들이 우리 집에 왔다.	[사람/NNG+들/XSN]
그들은 밥을 먹고들 싶었다.	[먹/VV+고/EC+들/XSN]

② '-님'은 다음과 같이 세 가지의 분석 중의성을 가지므로 주의해서 분석한다.

'임'의 의미로 쓰인 경우: 보통명사

[예시] 님과 이별하다.	[님/NNG+과/JKB]

사람의 '이름'이나 '성' 뒤에서 쓰인 경우: 의존명사

[예시] 김철수님께서 오셨습니다.	[김철수/NNP+님/NNB+께서/JKS]

그 밖의 경우: 명사파생접미사

[예시] 과장님이 부르십니다.	[과장/NNG+님/XSN+이/JKS]

다) 동사파생접미사(XSV)

동사파생접미사는 어기 또는 어근에 붙어서 그것을 동사로 만들어 주는 기능을 갖는 접미사이다. 여기서는 그러한 접미사 중 현재 생산성을 가지고 쓰이는 것만을 인정하여 분석한다. 접사를 분석하고 난 나머지 언어 단위가 '어근(XR)'일 경우에는 더 이상 분리하여 분석하지 않고 통합한다.

당하	아군이 공격당하는 데에는 이유가 있다.	[공격/NNG+당하/XSV+는/ETM]
되	아침식사가 이미 준비되어 있었다.	[준비/NNG+되/XSV+어/EC]
시키	오늘 강아지를 운동시키려고 공원에 나갔다.	[운동/NNG+시키/XSV+려고/EC]
하	외국에서 공부하는 일이 쉬운 것은 아니다.	[공부/NNG+하/XSV+는/ETM]
받	몇몇은 집세 인상을 강요받았다.	[강요/NNG+받/XSV+았/EP+다/EF+./SF]

라) 형용사파생접미사(XSA)

형용사파생접미사는 어기나 어근에 붙어서 그것을 형용사로 파생시키는 접미사이다. 여기서는 그러한 접미사 중 현재 생산성을 가지고 쓰이는 것만을 인정한다. 접사를 분석하고 난 나머지 언어 단위가 '어근(XR)'일 경우에는 더 이상 분리하여 분석하지 않고 통합한다.

답	사람이 사람답게 행동해야지.	[사람/NNG+답/XSA+게/EC]
되	거짓된 말은 들통나기 마련이다.	[거짓/NNG+되/XSA+ㄴ/ETM]
롭	어려운 일일수록 슬기롭게 대처하라.	[슬기/NNG+롭/XSA+게/EC]
스럽	그녀의 사랑스러운 표정을 보거라.	[사랑/NNG+스럽/XSA+ㄴ/ETM]
하	건강한 신체에 건강한 정신이 깃든다.	[건강/NNG+하/XSA+ㄴ/ETM]

주의사항

명사구에 결합하는 '만하'(예: 집채만 하다)는 '만'을 보조사로, '하'는 그 활용 양상을 참고하여 형용사로 분석한다. '만하'는 앞에 관형사형이 올 경우 '만/NNB+하/XSA'로 분석되는 경우도 있으므로 주의해야 한다.

[예시]	그 일을 처리하는 데 철수만한 인재가 없다.	[철수/NNP+만/JX+하/VA+ㄴ/ETM]
	이 음식은 먹을 만하다.	[먹/VV+을/ETM] [만/NNB+하/XSA+다/EF+./SF]

3) 어근(XR)

국어에는 하나의 단어가 조사에 의해 분리되는 현상이 있다. 즉, 파생된 용언에서 보조사 등

의 삽입에 의해 어근과 접사가 분리되는 현상이 있다. 어절 분석 표지에 어근(XR)이 포함되어 있으므로 분리된 어근에 어근 표지를 할당할 수 있다.

[예시] 따뜻도 하다	[따뜻/XR+도/JX] [하/VA+다/EF]

신문기사의 제목 등에서 어근으로 문장이 끝나는 경우에도 어근 표지를 할당할 수 있다.

[예시] 대회 3일 차 분위기 무난	[무난/XR]

주의사항

① 형용사의 어근 분리 시 어근 뒤에 오는 '하다'에는 형용사(VA) 표지를 부여한다. 실제 '하다'가 형용사라고 보기 어려운 측면이 있으나, 그 활용 양상을 참고하고 말뭉치에서의 활용을 생각했을 때에는 형용사로 파악하는 것이 실익이 있다. 아래와 같은 경우도 '하다'를 형용사(VA)로 본다.

[예시] 영희는 키가 철수만 하다.	[철수/NNP+만/JX] [하/VA+다/EF+./SF]
비가 올 듯도 하다.	[듯/NNB+도/JX] [하/VA+다/EF+./SF]

② 용언을 파생하는 접미사(하다, 되다 등)를 분리하고 남은 단위가 어근(XR)일 경우에는 더 이상 분리하지 않고 통합형으로 분석한다.

[예시] 듬직하다	[듬직하/VA+다/EF]
취하다	[취하/VV+다/EF]

사 기타

1) 기호

가) 마침표

마침표(.)의 경우, 문장 끝에서 쓰인 것을 SF로 분석한다. 따옴표 안 인용문의 문장 끝에서 쓰인 마침표도 마찬가지이다.

[예시]	아이가 듬직하다.	[듬직하/VA+다/EF+./SF]
	“아이가 착하다. 또 듬직하다.”라고 말했다.	[듬직하/VA+다/EF+./SF+”/SS+라고/JKQ]

아래와 같이 장, 절 등의 항목을 구분하기 위해 숫자 사이에, 또는 숫자 끝에 마침표가 쓰이는 경우가 있다. 이때 숫자 중간에 있는 마침표는 SP로, 숫자 끝에 있는 마침표는 SF로 분석한다.

[예시]	1. 서론	[1/SN+./SF]
	1.1. 연구 목적	[1/SN+./SP+1/SN+./SF]

소수점으로 쓰인 마침표, 날짜를 나타내는 숫자 사이에서 쓰인 마침표, 홈페이지 주소 속 마침표 등 문장 종결의 의미가 없는 마침표는 SP로 분석한다.

[예시]	3.14	[3/SN+./SP+14/SN]
	www.an.com	[www/SL+./SP+an/SL+./SP+com/SL]

말줄임표 대신 마침표가 쓰인 경우, 마침표의 개수에 관계없이 모두 묶어 SE로 처리한다.

[예시]	그리고….	[그리고/MAJ+…./SE]

나) 기타 기호(SW)

길이, 무게, 수효, 시간 따위의 수량을 수치로 나타내는 단위들 중 ‘미터, 그램, 리터’ 등은 의존명사(NNB)로, 외국어로 된 ‘m, g, l’ 등은 기호(SW)로 분석함에 유의한다. ‘제곱미터, 퍼센트포인트’ 등 사전에 한 단어로 올라 있는 단위 명사를 기호로 나타낸 것도 아래와 같이 하나의 기호로 분석한다.

[예시]	5㎡	[5/SN+㎡/SW]
	2%p	[2/SN+%p/SW]

한글이 원이나 괄호 속에 들어간 아래와 같은 기호는 SW로 처리한다.

[예시]	㈜/SW, ㈀/SW, ㉮/SW	
[참고]	(주)	[(/SS+주/NNG+)/SS]
	→ 이와 같이 괄호와 한글을 분리할 수 있는 경우에는 각각을 따로 분석한다.	

다) 한자(SH)

한자를 SH로 처리한다. 한자가 원이나 괄호 속에 들어간 기호도 SH로 처리한다.

[예시] ㊊/SH, ㈤/SH
[참고] (五) [(/SS+五/SH+)/SS]
→ 이와 같이 괄호와 한자를 분리할 수 있는 경우에는 각각을 따로 분석한다.

라) 외국어(SL)

한자(SH)를 제외한 외국 문자(로마자, 가나 등)를 SL로 처리한다. 로마자로 쓰인 숫자(로마숫자 I, II, III 등)도 로마자임에 주목하여 SL로 처리한다. 외국 문자가 원이나 괄호 속에 들어간 기호도 SL로 처리한다.

[예시] ⓐ/SL, Ⓐ/SL
[참고] (a) [(/SS+a/SL+)/SS]
→ 이와 같이 괄호와 외국 문자를 분리할 수 있는 경우에는 각각을 따로 분석한다.

아래와 같이 어떤 표현의 구체적인 내용을 숨기려는 의도로, 또는 구어 전사 시 말이 정확히 들리지 않아 로마자 X 표시를 사용하는 경우가 있다. 그런 경우 아래와 같이 처리한다.

[예시] 이런 버르장머리 없는 X [X/SL]
XXXX 이론 [XXXX/SL]
→ 이처럼 한 어절 전체가 X로 되어 있는 경우, 전체를 묶어 SL로 처리한다.
※ X 대신 O가 쓰일 때도 마찬가지이다.
※ 만약 로마자가 아닌 ×(곱셈표)나 △ 등이 쓰였으면 SW로 처리한다.
[예시] 어쩌라는 거야 씨X [씨X/NA]
XX스의 이론 [XX스/NF+의/JKG]
XXX의 이론 [XXX의/NA]
→ 이처럼 어절의 일부가 X로 되어 있는 경우, X를 포함하는 말의 품사를 고려하여 명사에 준하면 NF로, 용언에 준하면 NV로, 그 외에 해당하면 NA를 부여한다. 'XXX의'의 경우 XXX가 체언일 것으로 예상은 되지만 확실치 않으므로 전체 어절을 NA로 처리한다.

마) 숫자(SN)

아라비아숫자(0, 1, 2 등) 및 아라비아숫자가 원이나 괄호 속에 들어간 기호를 SN으로 처리한다.

[예시] ① 조리법	[①/SN]
⑴ 조리법	[⑴/SN]
[참고] ⑴ 조리법	[(/SS+1/SN+)/SS]
→ 이와 같이 괄호와 숫자를 분리할 수 있는 경우에는 각각을 따로 분석한다.	

바) 기호가 어절 중간에 개입한 경우

기호가 어절 중간에 개입한 경우, 기호를 뺀 말이 사전에 한 단어로 등재되어 있다면 기호가 있다 하더라도 전체를 통합하여 표지를 부여한다.

[예시] 농·수산물 (사전: 농수산-물)	[농·수산물/NNG]
초·중·고 (사전: 초중고)	[초·중·고/NNG]
의~리	[의~리/NNG]
사이~소 (사전: 어미 '-이소')	[사/VV+이~소/EF]

기호를 뺀 말이 사전에 한 단어로 등재되어 있지 않은 경우에도, 분리하여 분석할 경우 어근이 남는다면 전체를 통합하여 표지를 부여한다.

[예시] 당·정·청 (사전: 당정)	[당·정·청/NNG]

사전에 '당정'만이 등재되어 있어 이 어절을 '당·정/NNG'과 '·/SP', '청'으로 분리할 경우, 사전 미등재어이면서 홀로 쓰이지 않는 '청'이 남는다. 이 경우 '청'을 앞말에 통합하여 '당·정·청/NNG'로 표지를 부여한다.

단, 숫자나 외국어로만 표기된 경우에 기호가 포함되어 있으면 모두 각각 분석한다.

[예시] 6.25	[6/SN+./SP+25/SN]

아래와 같이 외국 문자나 숫자로 된 '주석'이 어절 중간에 개입하는 경우에는 각각의 요소를 분리하여 분석한다. 이 경우 표지를 줄 수 없는 불완전한 형태가 생길 수 있다.

[예시]마이크로소프트(microsoft)사	[마이크로소프트/NNP+(/SS+microsoft/SL+)/SS+사/NNG]

2) 준말

여러 개의 어미가 결합한 준말은 그 안에 분석 대상 선어말어미가 들어 있는 경우에 한해서만 복원한다.

[예시] 간다는	[가/VV+ㄴ다는/ETM] (○)
	[가/VV+ㄴ다/EF+하/VV+는/ETM] (×)
간댔어.	[가/VV+ㄴ다/EF+하/VV+았/EP+어/EF+./SF] (○)
	[가/VV+ㄴ댔어/EF+./SF] (×)

3) 분석불능범주

그 자체가 사전에 등재되어 있지도 않으면서, 축약의 정도가 심하거나 분석하기 어려운 방언형의 경우 분석불능범주로 처리한다. 분석이 어렵더라도 그 품사(범주)를 명확히 할 수 있는 경우에는, 추정 범주인 NF(명사 추정 범주), NV(동사 추정 범주)를 부여한다.

[예시] 담배가 쪼매턴게 하마 자라서 빼나?	[쪼매턴게/NA]

<우리말샘>에 접사로 등재되어 있으나 본 지침의 '분석 대상 접사'가 아닌 요소가 앞뒤의 기호 등 때문에 분리되어 홀로 남은 경우, 해당 요소를 분석불능범주로 처리한다. <우리말샘>에 등재되어 있지 않고 홀로 쓰이지도 않아 어근에 준하는 것으로 볼 수 있는 요소가 같은 이유로 홀로 남은 경우에도, 해당 요소를 분석불능범주로 처리한다.

[예시] 대(對)중국	[대/NA+(/SS+對/SH+)/SS+중국/NNP]
5인(人)승	[5/SN+인/NNG+(/SS+人/SH+)/SS+승/NA]

4) 합성어

<우리말샘>에 등재되어 있는 합성어를 한 단위로 둔다. 합성어가 북한어나 방언으로 등재되어 있어도 분석하려는 말과 의미가 동일하다면 표준어와 동일하게 처리한다.

주의사항

① <우리말샘>에 합성어로 올라 있는 단어는 한 단위로 분석한다.

[예시] 정치권력 (사전: 정치-권력)	[정치권력/NNG]

② 어절에 나타난 표기가 규범에 맞지 않아 사전에서 검색되지 않으나 규범에 맞게 표기된 단어는 사전 등재어일 때, 규범에 맞게 표기된 단어에 준하여 한 단어로 분석한다.

[예시] 먼저번 (사전: 먼젓번)	[먼저번/NNG]
조랭이떡 (사전: 조롱이-떡)	[조랭이떡/NNG]

③ 단어 자체가 사전의 표제어로 등록되어 있지는 않으나 사전에 구로 등재되어 있는 말(A^B)의 일부에 해당하는 단어일 때에도 한 단위로 분석한다.

[예시] 사대강을 (사전: 사대강^수계법)	[사대강/NNG+을/JKO]

④ 사전에 구로 등재되어 있는 말(A^B)은 세분하여 분석하는 것을 원칙으로 한다.

[예시] 학생운동 (사전표기: 학생^운동)	[학생/NNG+운동/NNG]

구를 이루는 둘 이상의 요소를 분리했을 때, 어느 한 요소에 분석 대상 접사가 포함되어 있는 경우가 있을 수 있다. 이 경우 물론 분석 대상 접사를 분리해야 한다.

[예시] 인적사항 (사전: 인적 사항)	[인/NNG+적/XSN+사항/NNG]

⑤ 아래와 같은 혼성어는 분리하여 분석하기 어려우므로 한 단어로 보고, 의미에 따라 일반명사 또는 고유명사로 분석한다.

[예시] 아베노믹스 (아베+이코노믹스) [아베노믹스/NNG]
→ 이는 아베의 경제정책을 일컫는 말로, 의미상 본 지침의 고유명사 부류에 들지 않는다. 이에 따라 전체 단어를 일반명사로 처리한다.
[예시] 홍드로 (홍수아+페드로) [홍드로/NNP]
→ 이는 특정 인물을 가리키는 말로 사용되고 있으므로 의미상 고유명사 부류에 든다. 이에 따라 전체 단어를 고유명사로 처리한다.

⑥ 합성어로 등록되어 있지 않은 표제어는 분리해서 분석하되, 사전 표제어로 등록되어 있는 최대한 많은 음절수의 단어를 생성하도록 나눈다. 즉 다음 예와 같은 경우 3음절 어휘가 생성되는 첫 번째 분석을 취한다.

[예시] 영상학과 [영상학/NNG+과/NNG] (3음절+1음절)
영상학과 [영상/NNG+학과/NNG] (2음절+2음절)

⑦ 3음절 어휘와 같이 어느 쪽으로 나뉘어도 음절수가 같고, 양쪽 분석이 모두 사전 표제어라면 뒤쪽을 먼저 분석한다.

[예시] 차창밖 [차/NNG+창밖/NNG]
이등품 [이/NR+등품/NNG]

⑧ 합성어로 등록되어 있지 않은 표제어를 더 작은 요소로 분리했을 때 어근이 남거나 품사를 부여하기 어려운 요소가 남는다면, 해당 요소를 분리하지 않고 앞말 또는 뒷말과 결합하여 형태 표지를 부여한다.

[예시] 당정청 [당정청/NNG]

사전에는 '당정청'이 올라 있지 않고 '당정'만이 올라 있다. 청와대를 뜻하는 '청'은 미등재어이고 홀로 쓰이는 일이 드물어 어근으로 판단할 수 있다. 이런 경우 '청'을 앞말인 '당정'과 결합하여 처리한다.

[예시] 오인승 [오인승/NNG]

'오인승'은 수사 '오'와 명사 '인', 그리고 사전 미등재어인 '승'으로 구성되어 있다. '승'은 미등재어이지만 홀로 쓰이지 않으므로 어근으로 판단할 수 있고, 이런 경우 '승'을 앞말인 '인'과 결합하여 처리한다. 그런데 그렇게 해서 도출된 '인승' 역시 미등재어이고, '인승'의

'인'이 일반명사임을 고려하면 '인승'도 일반명사가 되어야 할 것이나 이 말이 홀로 쓰이지 않기 때문에 일반명사로 품사를 부여하기가 어렵다. 따라서 '인승'도 분리하지 않고 앞말과 결합하여 형태 표지를 부여한다.

5) 파생어

<우리말샘>에 등재되어 있는 파생어를 한 단위로 둔다.

주의사항

① 사전에 파생어가 등재되어 있어도 그 안에 분석 대상 접사가 포함되어 있으면 분석한다. 접사의 분석 범위는 접사 지침의 주의사항에 따른다.

[예시] 수습생 (사전: 수습-생)	[수습/NNG+생/XSN]

② 사전에 구로 등재되어 있는 말 안에 분석 대상 접사가 포함되어 있는 경우 역시 접사를 분석한다. 단, 접사를 분리해 냈을 때 어근이 남는 경우에는 접사를 분리하지 않는다.

[예시] 도선수습생 (사전표기: 도선^수습생)	[도선/NNG+수습/NNG+생/XSN]

③ 분석 대상 접사를 분리한 후 남은 단위가 사전 미등재어인 경우가 있다. 해당 미등재어가 홀로 쓰이지도 않고 조사와도 결합하지 않는다고 판단된다면 그것을 어근으로 보아 접사를 분리하지 않는다. 단, 해당 미등재어가 어근이 중첩된 형식이라면 접사를 분리한다.

[예시] 가급적이면	[가급적/NNG+이/VCP+면/EC]
대대적 개편	[대대적/MMA]

'-적'이 분석 대상 접사인데 '가급'과 '대대'가 미등재어이다. '가급' 및 '대대'는 홀로 쓰이지도 않고 뒤에 조사가 결합할 수도 없는 단위이므로 미등재어이더라도 어근으로 보아 '가급적', '대대적'에서 '-적'을 분리하지 않는다. '가급적'은 사전에 명사와 부사로, '대대적'은 관형사와 명사로 등재되어 있으므로, 뒤에 조사가 후행하는 경우에는 명사로, 조사 없이 체언이 후행하는 경우에는 관형사로, 그렇지 않은 경우에는 부사로 맥락에 맞게 분석한다.

[예시] 나른나른한	[나른나른/MAG+하/XSA+ㄴ/ETM]

'-하-'가 분석 대상 접사인데 '나른나른'이 미등재어이다. '나른나른'은 사전에 등재된 어근 '나른'의 중첩형이다. 이 경우 '-하-'를 분리하고, '나른나른'에 일반 부사 표지를 부여한다.

④ 분석 대상 접사 목록에 없는 접사(비분석 접사)가 결합한 단어는, 그것이 사전 미등재어여도 한 단위로 둔다.

[예시] 임명자	[임명자/NNG]

'임명자'는 미등재어이고 '-자'는 접미사인데 분석 대상 접사는 아니다. 이 경우 '-자'를 '임명'에 결합하여 처리하지 않으면 달리 처리할 수 있는 방법이 없다. 따라서 '임명자'를 한 단위로 두고 일반명사 태그를 부여한다.

아 구어

구어 자료의 형태 분석 방법은 기본적으로 문어 자료의 형태 분석 방법과 동일하다. 다만 구어에서 나타나는 준말과 형태 변이 현상을 되도록 분석에 반영하고, 구어 전사 시 이용된 특별한 마크업과 표지를 처리하기 위해 아래의 지침을 별도로 마련하였다.

1) 구어에서 나타나는 준말과 형태 변이 현상의 처리

가) 하나의 요소 내부에서 형태 변이가 일어난 경우

아래와 같이 하나의 형태 표지가 붙는 단위에 구어의 음성적 특성이 반영되어 형태 변이가 일어났을 때는, **원어절의 형태를 바꾸지 않되 표준형에 비추어 형태 표지를 부여**한다.

[예시]	**건**(<그건) 어렵지 않아요	[**거/NP**+ㄴ/JX]
	것두(<그것도) 좋은데	[**것/NP**+두/JX]
	늦을까 봐 **날라**서 왔어.	[**날르/VV**+아서/EC]
	이**걸로**	[이거/NP+**ㄹ로/JKB**]
	좋**으까**?	[좋/VA+**으까/EF**+?/SF]
	여기 앉**어**	[앉/VV+**어/EF**]
	그렇게 하**더래도**	[하/VV+**더래도/EC**]
	학교 **간대더라**	[가/VV+**ㄴ대더라/EF**]
	할런지 모르겠다	[하/VV+**ㄹ런지/EF**]
	갈 것 같**애**	[같/VA+**애/EF**]

→ '애'가 이처럼 표기상으로 분리되어 드러난 경우에만 '애'로 분석한다. '가기를 바래', '나만 나무래'에서처럼 '애'가 표기상으로 분리되어 드러나지 않은 경우에는 모음조화에 따라 '아'로 분석한다.

[예시] 가기를 바**래** [바라/VV+**아/EF**]

나) 본래 둘 이상으로 분석되어야 하는 요소인데 축약되어 형태 분리가 어려워진 경우

(1) 용언 어간과 어미의 결합형인 경우

사전에 한 단어로 올라 있는 말이 아니어서 용언 어간과 어미로 분석해야 하는 말이 다음과 같이 축약되어 전사된 경우가 있다. 이때는 아래와 같이 구어의 변이 형태를 그대로 인정하면서 어간과 어미를 분리하여 분석한다. 형태상 분리가 어려움에도 어간과 어미를 분리하도록 한 것은, 절을 꾸리는 데 있어서 용언의 역할이 중요한 만큼 용언 어간의 모습을 드러낼 필요가 있기 때문이다.

(가) 형용사 '이렇-', '그렇-', '저렇-', '어떻-'류의 변이 형태

[예시]	일케	[잃/VA+게/EC]
	이케	[잏/VA+게/EC]
	이르케	[이룽/VA+게/EC]
	요러케	[요렇/VA+게/EC]
	요렇케	[요렇/VA+게/EC]

→ 어미가 '케'로 잘못 전사되었지만 'ㅎ'과 '게'가 만나 '케'가 된 것으로 보아야 하므로

'게/EC'로 분석한다.

요케 [욯/VA+게/EC]

요로케 [요롷/VA+게/EC]

그르케 [그룷/VA+게/EC]

그런케 [그럲/VA+게/EC]

그러치 [그러치/IC]

→ 사전에 '그렇지'가 "틀림없이 그렇다는 뜻으로 하는 말"로서 감탄사로 올라 있다. 그 의미로 쓰인 '그러치, 그룷치, 그치'는 더 분석하지 않고 감탄사로 보아야 한다.

→ '그렇지'로 전사되어야 할 것이 '그러치'로 전사되었다. 하지만 위에서 본 '요렇케'의 경우와 달리 내부 형태 분석을 하는 상황이 아니므로 원문의 표기를 그대로 따른다.

그룷치 [그룷치/IC]

그치 [그치/IC]

그치 않습니까? [긇/VA+지/EC]

그룷치 않습니까? [그룷/VA+지/EC]

→ 이때는 '그치', '그룷치'가 감탄사로 쓰인 것이 아니다. 따라서 위와 같이 용언 어간과 어미로 분석해야 한다.

→ '그룷치'는 '그룷지'로 전사되어야 할 것이 잘못 전사된 것이다. 잘못 전사된 부분에서 형태 분석이 이루어지므로, 위에서 본 '요렇케'의 경우와 마찬가지로 '치/EC'가 아니라 '지/EC'로 분석한다.

그찮아 [긇/VA+잖아/EF]

어뚱케 [어뚱/VA+게/EC]

→ '어뚱게'로 전사되어야 할 것이 잘못 전사된 것이다. 잘못 전사된 부분에서 형태 분석이 이루어지므로, 위에서 본 '요렇케, 그룷치'의 경우와 마찬가지로 '케/EC'가 아니라 '게/EC'로 분석한다.

어뜨케 [어뜩하/VV+아/EF]

→ '어떡해'의 변이형이다. 사전에 '어떡하다'가 등재되어 있어 '어떡해'를 '어떡하/VV+아/EF'로 분석하는 것을 참고하여 '어뜩하/VV+아/EF'로 분석한다.

→ 물론, '어뜨케 됐어?'에서처럼 부사어로 쓰인 것은 '어뚱/VA+게/EC'가 될 것이다.

(나) 그 외 용언의 변이 형태

[예시] 따르케 [따릉/VA+게/EC]
다르케 [다릉/VA+게/EC]
요만하케 [요만항/VA+게/EC]
→ '다르다', '요만하다'의 변이 형태가 나타났다. 역시 변이 형태를 그대로 인정하여 용언 어간과 어미를 분리한다. '따르+케', '요만하+케'로 분석될 수도 있을 것이나 가능한 한 용언 어간 쪽에서 변이 형태를 인정하기로 한다.

(다) 두 어절 이상에 해당하는 용언 어간+어미의 경우

[예시] 어케요 [엏/VA+게/EC+하/VV+아요/EF]
→ 두 어절에 해당하는 '어떻게 해요'가 줄어들었고, 그 속에 용언 어간과 어미가 있다. 용언 어간 '하'가 생략되었는데, 이때 '하'를 복원하지 않으면 어절 속에 동사의 어간이 없는 셈이 되므로 '하'를 복원해야 한다. '하' 앞의 '어케'는 '엏/VA+게/EC'로 분석한다.
이케서 [잏/VA+게/EC+하/VV+아서/EC]
→ 역시 두 어절에 해당하는 '이렇게 해서'가 줄어들었다. 용언 어간 '하'를 복원하지 않으면 어절 속에 동사의 어간이 없는 셈이 되므로 복원해야 한다. '하' 앞의 '이케'는 '잏/VA+게/EC'로 분석한다.
왜케 [왜케/MAG]
왤케 [왤케/MAG]
→ 두 어절에 해당하는 '왜 이렇게'가 줄어들었다. '왜케', '왤케'가 절에서 서술어로 쓰이는 일은 없으므로, 이 경우에는 예외적으로 더 분석하지 않고 '왜케, 왤케'를 일반부사로 처리한다.

(라) '이케~'와 같이 원문에 물결표 표시가 있는 것은 물결표를 제외하고 IC로 분석해야 함에 유의한다.

(마) 'X하-' 형태에서 '하'가 아예 생략되거나 'ㅎ'만 남은 아래와 같은 경우에는 '하'의 형태를 완전하게 복원한다. '하'를 복원하지 않으면 어절 속에 용언의 어간이 없는 셈이 되므로 복원하지 않을 수 없다.

[예시] 논의토록 [논의/NNG+하/XSV+도록/EC]
생각지 못한 [생각/NNG+하/XSV+지/EC]

(2) 용언 어간과 어미의 결합형이 아닌 경우

용언 어간과 어미의 결합형이 아니라면, 아래와 같이 형태 분리가 어려운 구어의 축약형을 더 분석하지 않고 하나의 단어로 인정하는 방안을 취하기로 한다. 형태 분리가 어려운 경우란, 형태 분리를 했을 때 적어도 한 요소가 사전 미등재어이고 그 요소가 다른 환경에서는 나타나지 않는 경우를 말한다. 형태 표지는 해당 단어의 문장 성분을 고려하여 부여한다(예: 부사어→부사).

[예시] 내비둬 [내비두/VV+어/EF]

→ '두-'는 분석 가능하지만 '내비'가 사전 미등재어이다. 그리고 '내비'는 '두-' 앞 외의 다른 환경에서는 거의 나타나지 않는다. 이에 따라 '내비두-' 전체를 동사로 처리한다.

냅둬 [냅두/VV+어/EF]

여따(<여기에다가) 놔. [여따/MAG]

→ '여'는 등재되어 있지만 '따'가 미등재어이다. 이 '따'는 '여따, 거따, 저따' 외에서는 보기 어렵다. 이에 따라 '여따' 전체를 한 단어로 처리한다. 문장 속에서 부사어로 쓰이므로 일반부사로 처리한다.

언놈이(<어느 놈이) 그래? [언놈/NP+이/JKS]

→ '놈'은 분석 가능하지만 '언'이 사전 미등재어이다. 그리고 '언'이 '놈' 앞 외의 다른 환경에서는 나타나지 않는다.

얻다 대고 [얻다/MAG]

→ '-다'는 분석 가능하지만 '얻'이 사전 미등재어이다. 그리고 '얻'이 '-다' 앞 외의 다른 환경에서는 나타나지 않는다.

클났다. [클나/VV+았/EP+다/EF+./SF]

→ '클'이 사전 미등재어이고 '나다' 외의 다른 환경에서 나타나지 않는다.

어서(<어디서) 그래? [어서/MAG]

→ '어'가 사전 미등재어이고 '서' 외의 다른 환경에서 나타나지 않는다.

짱난다 [짱나/VV+ㄴ다/EF]

→ '짱'이 사전 미등재어이고 '나다' 외의 환경에서 나타나지 않는다.

주의사항

① ‘이리로, 그리로, 저리로, 요리로, 고리로, 조리로’뿐 아니라 ‘일로, 글로, 절로, 욜로, 골로, 졸로’가 <우리말샘>에 부사로 등재되어 있으므로 MAG로 분석해야 함에 유의한다.

② 아래와 같이 같은 모음이 겹치면서 축약된 경우에는 본래 형태를 복원한다.

[예시] 어딨어요.	[어디/NP+있/VA+어요/EF+./SF]

다) 비표준적인 준말 활용형

아래와 같이 비표준적인 준말 활용형이 나타난 경우, 용언 어간은 표준형으로 복원하되 어미에서 매개모음 ‘으’를 빼고 분석한다.

[예시] 여기다 논(<놓은) 거야.	[놓/VV+ㄴ/ETM]
여기다 노셨던(<놓으셨던) 거야.	[놓/VV+시/EP+었/EP+던/ETM]
찌시더니(<찧으시더니)	[찧/VV+시/EP+더니/EC]
아이를 나면은(<낳으면은)	[낳/VV+면/EC+은/JX]

라) 삼중모음

모음 ‘ㅟ’와 ‘ㅓ’가 축약되어 삼중모음 발음이 나타난 경우, ‘사귀’어요’와 같이 ’로 표시되어 있다. 형태 분석 시에 이 ’는 반영하지 않는다.

[예시] 바뀌’었었어요.	[바뀌/VV+었었/EP+어요/EF+./SF]

마) 관형격조사 ‘에’

관형격조사 ‘의’의 발음을 ‘에’로 전사한 경우가 있다. 이 경우 ‘에/JKG’로 형태 표지를 붙인다. ‘에’가 부사격조사인지 관형격조사인지 판단이 어려운 경우에는 부사격조사(JKB)로 형태 표지를 붙인다.

[예시] 나**에** 생각	[나/NP+**에/JKG**]

바) 지정사 '이다'

구어에서 이중모음이 단모음으로 발음되는 현상이 자주 일어나 그 결과 '예'가 '에'로 발음되고, 아래와 같이 지정사 '이다'가 생략된 것으로 보이는 현상이 있다. 이때는 문법적으로 지정사가 있으나 단지 이중모음이 단모음으로 발음된 것으로 보아 지정사를 복원한다.

[예시] 이렇게 얘기할 거에요.	[거/NNB+**이/VCP**+에요/EF+./SF]

사) 구어에서 나타나는 사전 미등재 요소

사전에 등재되지 않은 문법 요소 '-ㄹ랑', '-ㄹ동'은 사전에 등재된 연결어미 '-ㄹ락'을 참고하여 연결어미로 분석한다.

[예시] 이해가 갈랑말랑 하길래	[가/VV+**ㄹ랑/EC**+말/VV+**ㄹ랑/EC**]

의성의태어를 구성하는 요소가 여러 번 반복되어 나오는 경우의 처리는 아래와 같다.

[예시] 지글지글지글지글 [지글지글/MAG+지글지글/MAG]

→ 사전에는 '지글지글'이 한 단어로 올라 있고, '지글'은 어근에 해당한다. 위의 경우 단어 '지글지글'이 두 번 연달아 나온 것으로 분석할 수 있으므로 두 단어로 나누어 처리한다.

지글지글지글 [지글지글지글/MAG]

→ '지글지글'을 한 단어로 처리하면 어근에 해당하는 '지글'이 남는다. 이런 경우에는 어근을 앞말에 붙여서 '지글지글지글' 전체를 하나의 단어로 처리한다.

지글지글지글지글지글 [지글지글/MAG+지글지글지글/MAG]

→ 위의 경우에는 뒤쪽에 더 많은 음절수가 남도록 지글지글/MAG+지글지글지글/MAG로 분석한다.

2) 구어 전사 시 이용된 마크업과 표지의 처리

가) 물결표(~)

머뭇거림을 나타내는 담화표지에 ~(물결표)가 붙어 있다. 이 경우 ~는 분석에서 제외하고 ~ 앞에 있는 말에 IC를 부여한다.

[예시] 아~	[아/IC]
그~	[그/IC]
뭐~	[뭐/IC]

단, 아래와 같이 머뭇거림을 나타내는 담화표지가 아닌 것에 ~(물결표)가 붙어 있는 경우가 있다. 그런 경우는 물결표를 넣지 않아야 할 곳에 넣은 전사 오류에 해당하므로, 물결표를 분석에서 제외하고, 남은 요소에 형태 표지를 부여한다.

[예시] 국호를~을~	[국호/NNG+를/JKO+을/JKO]

나) 마크업 기호

<trunc>, </trunc> 등의 마크업 기호는 한 어절로 두고 형태 표지를 부여하지 않는다. 단, 아래의 주의사항에 유의한다.

주의사항

① <note> </note> 마크업의 경우에는 마크업 기호와 그 안의 내용을 모두 한 줄에 보여주고, 어떠한 표지도 부여하지 않는다.

[예시] <note>배경 화면 잠깐 나옴</note> → 한 어절로 두고 분석하지 않음.

② 사람 이름, 주소 등 개인 정보 보호를 위한 마크업은 다음과 같은 방식으로 형태 표지를 부여한다.

[예시] <anon type="**name**" n="**1**"/>가	[**name1/NNP**+가/JKS]
<anon type="**name**"/>가	[**name/NNP**+가/JKS]
<anon type="**address**" n="**2**"/>은	[**address2/NNP**+은/JX]

구어 전사 지침상, 일반 대화에서 대화자들 및 관련인의 개인 정보가 드러난 경우에는 개인 정보 보호를 위하여 해당 정보를 위와 같이 마크업으로 가리도록 하였다. 그런데 전사 실수로 이름 등의 개인 정보가 마크업 없이 노출된 경우가 있다. 이때에는 해당 개인 정보에 NAP 표지를 부여하고, 향후 보완 방법을 모색할 수 있도록 한다.

NAP 표지는 일반 대화 자료의 형태 분석에서만 적용하며, 공적 방송 자료의 형태 분석

에서는 적용하지 않는다. 또한 일반 대화 자료에서도 정치인, 연예인 등 유명인의 이름에는 적용하지 않는다.

[예시] 지현이가 그러는데 [**지현이/NAP**+가/JKS]

③ 사람 이름에서 분리된 접미사 '-이'에는 NA를 부여한다.

[예시] <anon type="name" n="5"/>**이** 말고 하나가 더 있니?
→ '말고' 앞에는 주격조사가 올 수 없다. 이때 '말고' 앞에 나온 '이'는 '영숙이'에서 볼 수 있는 접사 '이'로 판단 가능하다. 본 지침의 비분석 접사가 분리되어 나온 경우이므로 이러한 '이'는 '이/NA'로 처리한다.

다) <trunc> </trunc> 사이의 요소

<trunc> </trunc> 사이에 표시되어 있는 끊어진 어절(단어가 불완전하게 발화된 경우)에는 NA(분석불능범주) 표지를 부여한다.

[예시] 미국과 <trunc>같</trunc> 같은
→ <trunc>
같 [같/NA]
</trunc>

단, <trunc> </trunc> 사이에 있는 요소를 제외할 경우 앞뒤의 말이 이어지지 않는 경우라면, <trunc> </trunc> 사이에 있는 요소이더라도 형태 표지를 부여한다.

[예시] 또 <trunc>문의하</trunc> 하기도 했습니다.
→ <trunc>
문의하 [문의/NNG+하/XSV]
</trunc>
하기도 [하/XSV+기/ETN+도/JX]

라) <unclear> </unclear> 사이의 요소

전사 시 잘 들리지 않은 부분은 <unclear> </unclear>로 표시되어 있다. 가능한 한 각 요소에 맞는 형태 표지를 부여하고, 형태 표지를 부여할 수 없는 경우에는 NA(분석불능범주),

NV(용언추정범주), NF(명사추정범주)를 부여한다.

(1) 정확히 들리지 않았으나 x 표시 없이 전사된 경우

가능한 한 각 요소에 맞는 형태 표지를 부여한다.

[예시] <unclear>더 힘들어</unclear>
→ <unclear>
더 [더/MAG]
힘들어 [힘들/VA+어/EF]
</unclear>
[예시] 있<unclear>어요</unclear>
→ 있 [있/VA]
<unclear>
어요 [어요/EF]
</unclear>
[예시] 있어<unclear>요</unclear>
→ 있어 [있/VA+어/EF]
<unclear>
요 [요/JX]
</unclear>

(2) 일부 음절이 들리지 않은 경우

일부 음절이 들리지 않은 경우에는 해당 음절이 x로 표시되어 있다. 이때는 x가 포함된 단어 부분에 NA(분석불능범주), NF(명사추정범주), NV(용언추정범주)를 부여한다.

[예시] <unclear>xx스의</unclear> 이론을
→ <unclear>
xx스의 [xx스/NF+의/JKG]
</unclear>
이론을 [이론/NNG+을/JKO]

(3) <unclear> </unclear> 마크업으로 인해 표지를 주기 어려운 음절이 발생하는 경우

아래와 같이 <unclear> 마크업으로 인해 단어가 분리되어 표지를 주기 어려운 음절이 발생하는 경우에는, 각 음절에 NA(분석불능범주)를 부여한다.

[예시] 임시정 <unclear>부</unclear>		
→	임시정	[임시/NNG+정/NA]
	<unclear>	
	부	[부/NA]
	</unclear>	

3) 전사 오류 및 해석 불능 어절의 처리

가) 탈자로 인해 형태 표지 부여가 어려운 경우

아래와 같이 전사 과정에서 탈자가 발생하였거나 혹은 발화 실수로 과도한 생략이 일어나 형태 표지 부여가 어려워지는 경우가 있다. 이런 경우에는 해당 요소에 NA(분석 불능 범주), NV(용언 추정 범주), NF(명사 추정 범주) 중 하나를 부여한다.

[예시] 하지 못는(<못하는)	[못/NV+는/ETM]

나) 잉여적인 요소가 덧붙은 경우

아래와 같이 전사 과정에서 첨자가 발생하였거나 혹은 발화 실수로 잉여적인 형태가 덧붙은 경우가 있다. 이런 경우에는 해당 요소에 최대한 그 요소에 맞는 형태 표지를 부여하고, 만약 형태 표지 부여가 어렵다면 NA(분석 불능 범주)를 부여한다.

[예시] 국호를을	[국호/NNG+를/JKO+을/JKO]
됐습니다다.	[되/VV+었/EP+습니다/EF+다/EF+./SF]

다) 띄어쓰기 오류로 인해 형태 표지 부여가 어려운 경우

‘그럴걸’이 ‘그럴 걸’로, ‘뒤치락거리다’가 ‘뒤치락 거리다’로 띄어쓰기와 함께 전사된 경우가 있다. 이때 띄어쓰기 오류로 인해 ‘걸’의 처리, ‘거리-’의 처리가 어려워진다. ‘거리-’같이 용언의

성격을 띠는 요소에 대해서는 최대한 형태 표지를 부여한다. 그 외의 경우에도 최대한 형태 표지를 부여하지만, 형태 표지 부여가 어려운 요소에는 NA(분석불능범주), NV(용언추정범주), NF(명사추정범주)를 부여한다.

[예시] 뒤치락 거리고 [뒤치락/XR, 거리/VV+고/EC]
→ '거리-'는 본 지침에서 분석하지 않는 동사 파생 접미사이다. 용언의 성격을 띠는 요소에는 최대한 형태 표지를 부여하여 동사로 처리한다.

[예시] 그럴 걸 [그러/VV+ㄹ/ETM, 걸/NA]
→ 어미 '-ㄹ걸'이 분리되어 나왔는데, 앞의 '-ㄹ'에는 관형형 어미 표지를 줄 수 있지만 뒤의 '걸'은 처리가 어려우므로 분석 불능 표지를 준다.

[예시] 집에 갔는 지 모르겠다 [가/VV+았/EP+는/ETM, 지/NA]
→ 어미 '-는지'가 분리되어 전사되었는데, 앞의 '-는'에는 관형형 어미 표지를 줄 수 있지만 뒤의 '지'는 처리가 어려우므로 분석 불능 표지를 준다.

라) 표기법 오류로 인해 형태 표지 부여가 어려운 경우

아래와 같이 더 분석되어야 할 대상이 있음에도 표기법 오류 때문에 형태 분리 및 형태 표지 부여가 어려워지는 경우가 있다. 이 경우에는 올바른 표기법으로 수정한 형식을 상정하고 표지를 부여한다.

[예시] 공부를 하며는 [하/VV+면/EC+은/JX]
→ '면은'으로 써야 할 것을 '며는'으로 잘못 전사하였으며, 그 때문에 형태 분리가 어렵게 되었다. 이런 경우에는 올바른 표기형인 '면은'으로 복원하여 '면/EC+은/JX'로 형태 표지를 부여한다.

[예시] 너 때매 [땜/NNB+에/JKB]
→ '땜에'로 써야 할 것을 '때매'로 잘못 전사하였으며, 그 때문에 형태 분리가 어렵게 되었다. 이런 경우에는 올바른 표기형인 '땜에'로 복원하여 '땜/NNB+에/JKB'로 형태 표지를 부여한다.

[예시] 편찬되서 [편찬/NNG+되/XSV+어서/EC]
→ '돼서'로 써야 할 것을 '되서'로 잘못 전사하였다. 하지만 '되'와 '돼'는 발음이 동일하므로 단순한 표기법 차이 때문에 '어서/EC' 대신 '서/EC'로 형태 표지를 부여하는 것은 합리적이지 않다. 따라서 이 경우에도 올바른 표기형인 '편찬돼서'를 상정하여 '편찬/NNG+되/XSV+어서/EC'로 형태 표지를 부여한다.

아래와 같이 표기법 오류가 나타났으나 그 때문에 형태 분리 및 형태 표지 부여가 어려워지

는 상황이 아니라면, 원문의 형식을 그대로 두고 표지를 부여한다.

[예시] 크리마트를 했어요. [크리마트/NNG+를/JKO]

→ '크림아트'로 써야 할 것을 '크리마트'로 잘못 전사했다. 하지만 그 때문에 형태 표지 부여가 어려운 상황은 아니다('아트'가 미등재어이기 때문에 외국어 지침에 따라 '크림아트' 전체를 일반명사로 분석해야 하는 상황임). 이런 경우에는 원문의 표기를 그대로 두고 '크리마트/NNG'로 형태 표지를 부여한다.

[예시] 그렇다고 봅니다만은 [보/VV+ㅂ니다만은/EC]

→ '봅니다마는'으로 써야 할 것을 '봅니다만은'으로 잘못 전사했다. 하지만 그 때문에 형태 표지 부여가 어려운 상황은 아니다('-다마는'이 하나의 어미로 등재되어 있고, 그 앞에 선어말어미에 준하는 '습니'가 결합한 것으로 보아 '-습니다마는'을 하나의 어미로 분석해야 하는 상황임). 이런 경우에는 원문의 표기를 그대로 두고 'ㅂ니다만은/EC'로 형태 표지를 부여한다.

※ '숙제를 하긴 했다마는(했다만).'에서처럼 '-다마는'이 종결부에서 쓰일 때에도 뒤에 생략된 말이 있는 것으로 보고 연결어미 표지를 부여한다.

[예시] 얼마나 슬펐는 줄 아녜. [알/VV+녜/EF+./SF]

→ '냐'로 써야 할 것을 '녜'로 잘못 전사했다. 하지만 그 때문에 형태 표지 부여가 어려운 상황은 아니다('냐'는 '냐 해'의 준말로서 이처럼 '하'가 축약된 구성의 경우 그 속에 분석 대상 선어말어미가 들어 있지 않은 한 더 분리하지 않는다는 지침을 적용하여 '냐/EF'로 분석해야 하는 상황임). 이런 경우에는 원문의 표기를 그대로 두고 '녜/EF'로 형태 표지를 부여한다.

마) 어절의 의미 파악이 어려운 경우

어절의 의미를 파악하기 어려운 경우, 의미는 불분명하더라도 아래와 같이 해당 어절을 이루는 요소의 문법적 지위를 확정할 수 있다면 그에 따라 최대한 형태 표지를 부여한다. 문법적 지위를 확정하기 어렵거나 형태 표지를 부여하기 어려운 경우에 한해 NA를 부여한다.

[예시] 조선 웅 어~ 왕조 실의 역대 왕들의 왕릉 [실/NA+의/JKG]

→ 이때 '실의'의 의미를 정확히 파악하기가 어려우나, 맥락상 '조선 왕실의'에서 '실의'가 분리된 것으로 보인다. 이에 따라 실/NA+의/JKG로 형태 표지를 부여한다.

[예시] 아직까지 충청에 민심이 복마진입니다. [복마진/NNG+이/VCP+ㅂ니다/EF+./SF]

→ 이때 '복마진'의 의미를 정확히 파악하기가 어려우나 지정사 앞에 나타나는 문법적 특성으로 보아 일반명사로 판단할 수 있다. 이에 따라 복마진/NNG로 표지를 부여한다.

→ 위의 두 방식을 적용하여도 문법적 지위를 확정하기 어려운 요소가 있다면 해당 요소에 NA를 부여한다.

바) 한 어절 내에 마크업이 포함된 경우

다음과 같이 한 어절 내에 마크업이 포함된 경우에는, 마크업을 제외하고 남은 부분만을 지침에 따라 분석한다.

[예시] <trunc>아쉬</trunc>아쉬움이	[아쉬/NA+아쉬움/NNG+이/JKS]

제4장 결론

이 사업은 인공지능 발전을 위한 우리말 기초 자원으로 활용될 고품질의 한국어 형태 분석 말뭉치를 구축하고, 형태 분석 말뭉치 구축을 위한 표준적인 지침을 개발하는 데 주요 목적이 있다.

사업의 범위는 크게 두 부분으로 나눌 수 있다. 첫째는 형태 분석 말뭉치 구축 지침 수립으로, <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 언어 현실에 맞게 수정하고 보다 구체화하였다. 둘째는 형태 분석 말뭉치 구축으로, 형태 분석 말뭉치 구축 지침을 바탕으로 총 300만 어절 규모(문어 200만 어절, 구어 100만 어절)의 형태 분석 말뭉치를 구축하였다.

○ 형태 분석 말뭉치 구축 지침 수립

형태 분석 말뭉치 구축 지침을 수립하기 위하여 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 기반으로 삼되 문제점을 보완하였다. 이 과정에서 한국정보통신기술협회(TTA)의 '표준 형태소 태그셋(TTAK.KO-11.0010/R1)', <물결21>의 '형태 분석 지침'을 참조하였고, 형태 분석 질의응답 게시판을 운영하여 자주 질문되는 문제를 해결할 수 있는 설명과 사례를 지침에 추가하였다. 지침 보완의 방향성과 주요 보완 사항은 다음과 같다.

① 분석 지침의 구체화

- 조사 결합형을 분석하는 기준을 명시하였다.
- 접두사, 접미사를 언제 분리하는지에 대한 지침을 명시하였다.
- 언중의 직관을 고려하여 고유명사의 범위를 보다 넓히면서도, 작업의 일관성을

위하여 구체적으로 범위를 명시하고 다양한 사례를 제시하였다.

② 언어학적 엄밀성의 추구

- '있다'의 동사 용법과 형용사 용법을 구분하는 기준을 제시하였다.
- 종결어미와 연결어미를 후행하는 문장 부호에 따라 구분하기보다 기능에 따라 구분하도록 하였다.

③ 과도한 분석 지양

- 한국어에서 유의미하게 사용되는 언어 단위를 분석 대상으로 삼기 위하여 외국어 처리 지침과 기타 기호가 포함된 어절의 처리 지침을 재정비하였다.

또한 본 사업에서 마련한 지침은 문어뿐 아니라 구어 분석 시에도 적용할 수 있도록 구성되었다. 즉 문어와 구어는 원칙적으로 동일한 형태 표지 목록을 바탕으로 동일한 방식으로 분석된다. 다만 구어는 문어와 달리 다양한 준말과 형태 변이 현상을 보여 주므로 구어에서 나타나는 준말과 형태 변이 현상의 처리 방법을 따로 명시할 필요가 있다. 또한 구어 전사 시에 이용된 특별한 마크업과 표지가 있기에, 그것의 처리 방법도 별도로 제시할 필요가 있다. 이에 지침의 말미에 구어 분석 시의 유의점에 대한 지침을 붙였는데, 구어에서 나타나는 준말과 형태 변이 현상을 되도록 분석에 반영하는 방향으로 지침의 내용을 마련하였다.

○ 300만 어절 규모의 형태 분석 말뭉치 구축

이 사업에서 구축한 형태 분석 말뭉치의 총 규모는 300만 어절(문어 200만 어절, 구어 100만 어절)이다. 형태 분석 말뭉치의 기반이 된 원시 말뭉치는 2004년 이후 생산된 현대 국어 자료를 수집한 <2018년 국어 말뭉치 연구 및 구축> 사업 결과물의 일부로서, 문어 말뭉치는 전체 신문으로 구성되어 있으며 구어 말뭉치는 공적 독백, 공적 대화, 사적 대화를 포함한다. 원시 말뭉치의 각 어절을 대상으로 형태를 분리하고 형태 분류 표지(세분류 47종)를 부착하는 작업을 하였는데, 형태 분리의 기준이 되는 단위는 기본적

으로 <우리말샘>에 등재된 단어이되, 생산성이 비교적 높은 접사도 분리하는 것을 원칙으로 삼았다.

형태 분석 말뭉치 구축은 형태 분석 지침 수립 → 분석 도구(워크벤치) 구현 → 작업자 교육 → 자동 형태소 분석 → 분석 오류 수정 → 최종 결과물 산출의 순으로 이루어졌다.

이 중 분석 오류 수정은 3단계로 이루어졌다. 1단계는 작업자 2인이 동일한 원시 말뭉치에 대한 자동 형태 분석 결과를 각자 수정하는 단계이다. 2단계는 작업자 2인의 오류 수정 결과를 비교하며 검수자 1인이 형태 분석 결과를 검수하는 단계이다. 3단계는 전체 작업 결과물에 대해 상위 작업자 그룹이 형태 결합 오류 목록, 어절 분석 중의성 목록 등을 검토하며 오류를 수정하는 단계이다.

본 사업에서는 말뭉치 구축의 편의를 도모하고 정확성을 높이기 위하여 높은 분석 정확률을 갖춘 형태소 분석기(서울대 형태소 분석기)를 사용하였다. 서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다.

한편으로 형태 분석 말뭉치 구축에 최적화된 워크벤치를 개발하였다. 워크벤치에서는 서울대 형태소 분석기의 어절 분석 결과를 보여 주되 그것을 손쉽게 수정할 수 있게 하였고, 드롭다운 선택 방식 및 오류 검사를 통해 입력 오류를 원천적으로 차단하였다. 또한 동일 어절에 대한 작업자 2인의 분석 결과를 비교하고 분석 결과가 일치하지 않는 경우 그 중 옳은 분석을 선택할 수 있도록 하여 분석 결과 검수의 효율을 높였다.

Abstract

Conducting Korean POS tagged corpus

This project aims to build a high-quality Korean POS tagged corpus which is to be used as a basic resource for the development of artificial intelligence and develop standard guidelines for building a Korean POS tagged corpus.

The scope of this project can be largely divided into two parts. The first is to establish guidelines for building Korean POS tagged corpus, modifying and elaborating the guidelines from the 21st Sejong Project to bridge its gap from the reality of language use. Second is establishing Korean POS tagged corpus, covering a total of 3 million words (2 million written, 1 million spoken words) while following the guidelines for building Korean POS tagged corpus.

○Establishing guidelines for building Korean POS tagged corpus
To establish guidelines for building Korean POS tagged corpus, we depended on the guidelines from the 21st Sejong Project while fixing its errors. In this process, we referenced to 'TTAK.KO-11.0010/R1' from TTA(Telecommunications Technology Association), and 'guidelines for building Korean POS tagged corpus' from trends 21, and ran a QnA board for POS tagging to add solutions and examples for frequently asked questions to the guidelines. The direction and major complementation of the guidelines are as the following.

① To specify tagging guidelines
- Specified the standards for analyzing particle combinations.
- Specified the guidelines on when to separate prefixes and suffixes.

② To pursue linguistic precision

- Provided standards for distinguishing whether '*it-tɨ*' is used as a verb or as an adjective.
- Reorganized guidelines to distinguish between connective endings and final endings according to its function rather than the following punctuation.

③ To refrain from excessive analysis

- to make the significantly used language units in the Korean language the subject of analysis, we reorganized guidelines for foreign language processing and word-phrases containing other symbols.

Also, the guidelines provided in this project are intended to be applied when analyzing both written and spoken language. In other words, written and spoken language are basically analyzed in the same way. However, spoken language, unlike written language features diverse abbreviations and shape-shifting phenomena, which makes it necessary to provide methods to process them. Furthermore, because there are special markups and signs when transcribing spoken language, methods to process these must also be separately provided. Therefore, we attached a guideline of the considerations to make when analyzing spoken language, to put abbreviations and shape-shifting phenomenon into account, at the end of the guideline.

○ Building POS tagged corpus of 3 million words

This project covered a total of 3 million words (2 million written, 1 million spoken). The corpus which became the basis of the POS tagged corpus is a part of the outcomes form the project "2018 research and development of Korean corpus" which collected modern Korean language resources produced after 2004. The written corpus consists of only newspapers, and the spoken corpus consists of public monologue, public dialogue, and private conversations. We separated forms of each word and

attached a classification label (total 47 classes). The standard unit of separating forms were words enlisted in <Wurimalsaem>, while relatively productive affixes were separated as well.

POS tagged corpus was built in the following order; Building POS tagging guideline → implementing analyzer (workbench) → education → auto POS tagging → fixing errors → results

The stage of fixing errors was done in three stages. In step 1, two operators correct the same automatic POS tagging results of the raw corpus. In step 2, the two operators compare the results of error correction while one operator checks the results of form analysis. In step 3, the higher operator group reviews the list of errors in form combinations, ambiguity in word analysis and more of the whole output to fix errors.

In this project, we used a morphology analyzer with high accuracy (SNU morphology analyzer) to ease the process of building a corpus and enhance its accuracy. SNU morphology analyzer was developed by using the thoroughly fixed results of the Sejong morpheme-sense tagged corpus as its deep learning training set.

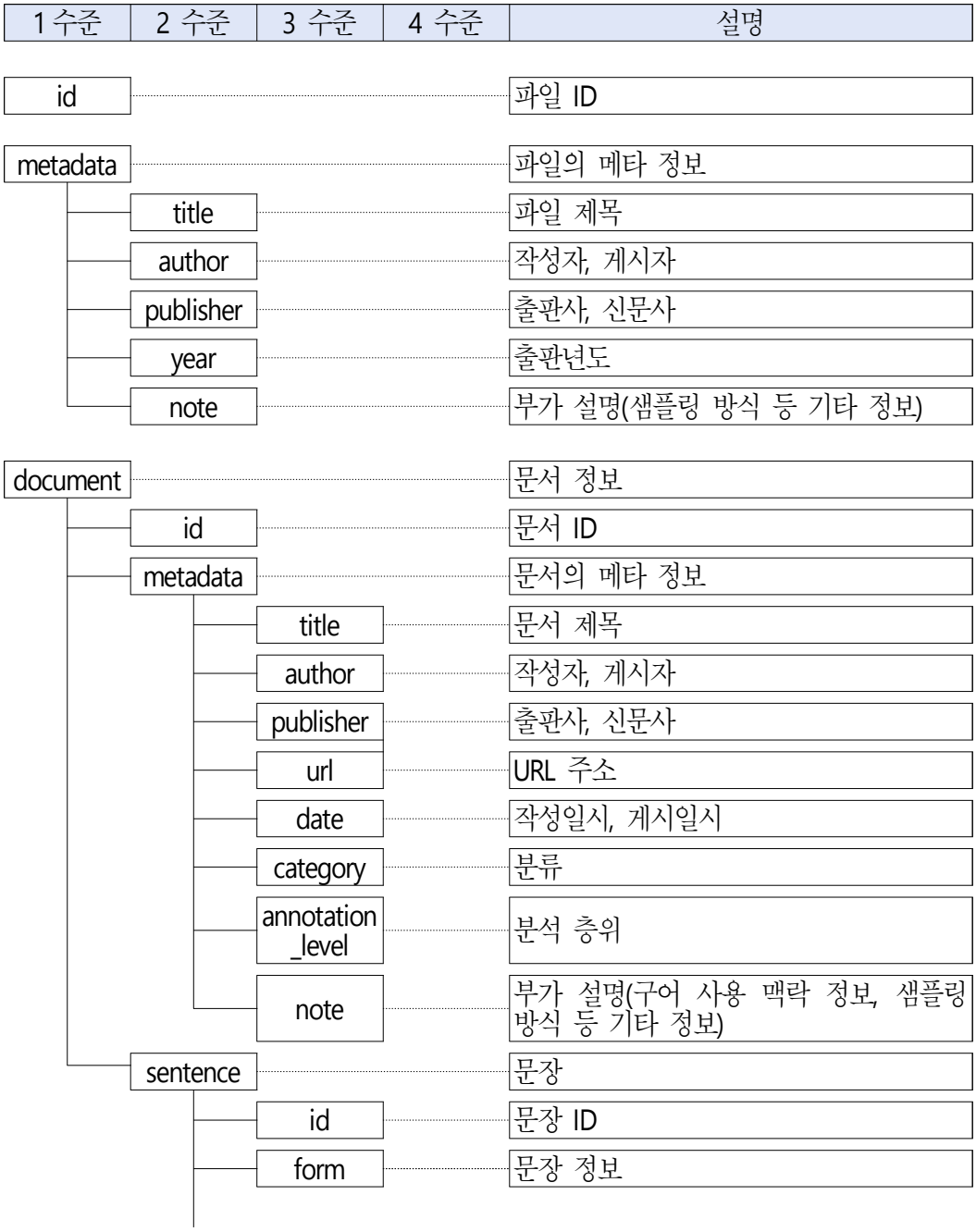
Meanwhile, we developed a workbench fit for building a POS tagged corpus. The workbench displayed the SNU POS taggers' word analysis results and made it easier to modify, using dropdown methods and error check to cut off input errors. Moreover, if the analysis results of the two operators about the same word did not match, it was possible to choose the correct one which made it possible to more effectively review the results.

연구 기관: 성신여자대학교 산학협력단 · 서울대학교 산학협력단 · 고려대학교 산학협력단

연구진	
연구 책임자	김일환(성신여자대학교)
공동 연구원	박진호(서울대학교)
	송상헌(고려대학교)
	유현조(서울대학교)
	윤태진(성신여자대학교)
	이규범(고려대학교)
	이도길(고려대학교)
	정성훈(목포대학교)
	정연주(홍익대학교)
	최운호(목포대학교)
연구 보조원	강규영(서울대학교)
	강수연(고려대학교)
	강주은(서울대학교)
	고동현(서울대학교)
	김동은(서울대학교)
	김연우(고려대학교)
	김연희(경희대학교)
	김영규(서울대학교)
	김은진(서울대학교)
	박종우(성신여자대학교)
	배윤정(서울대학교)
	서신애(성신여자대학교)
	윤성혁(서울대학교)
	이강혁(서울대학교)

	이세은(고려대학교)
	이영경(고려대학교)
	이용규(서울대학교)
	이윤복(서울대학교)
	이종진(성신여자대학교)
	전진호(서울대학교)
	정슬아(성신여자대학교)
	최선지(고려대학교)
	홍승혜(고려대학교)
	황현동(서울대학교)
	후박문(서울대학교)

부록 1: JSON 형식의 기본 구조




word
어절 정보
id
어절 ID
form
어절
begin
어절의 문장 내 시작 위치(UTF-8 문자 위치로 0 부터 시작)
end
어절의 문장 내 끝 위치(UTF-8 문자 위치로 0 부터 시작)
morpheme
형태소 분석 정보
id
형태소 ID
form
형태소
label
형태소 품사 태그
word_id
형태소가 포함된 어절 ID
position
어절 내 형태소 위치

부록 2: JSON 형식의 예시


```
{     "id": "WRAK00021",
      "metadata": {
            "title": "문어 200만 배포용",
            "author": "홍길동",
            "publisher": "국립국어원",
            "year": "2019",
            "note": "부분 추출 - 임의추출"
      },
      "document": [
      {
            "id": "WRAK0021.1",
            "metadata": {
                  "title": "줄다리기",
                  "author": "김철수",
                  "publisher": "국어일보",
                  "url": "https://www.korean.go.kr/",
                  "date": "20190709",
                  "category": "신문 > 전국 종합지",
                  "annotation_level": "형태 분석",
                  "note": "문서 부가 설명"
            },
            "sentence": [
                  {
                        "id": "WRAK0021.1.1",
                        "form": "줄다리기는 1900년 제2회 파리올림픽부터 1920년 제7회 앤트워프올림
                              픽까지 정식 종목이었다.",
                        "word": [
                              {"id": 1, "form": "줄다리기는", "begin": 0, "end": 5},
                              {"id": 2, "form": "1900년", "begin": 6, "end": 11},
                              {"id": 3, "form": "제2회", "begin": 12, "end": 15},
                              {"id": 4, "form": "파리올림픽부터", "begin": 16, "end": 23},
                              {"id": 5, "form": "1920년", "begin": 24, "end": 29},
                              {"id": 6, "form": "제7회", "begin": 30, "end": 33},
```

```
                    {"id": 7, "form": "앤트워프올림픽까지", "begin": 34, "end": 43},
                    {"id": 8, "form": "정식", "begin": 44, "end": 46},
                    {"id": 9, "form": "종목이었다.", "begin": 47, "end": 53}
                ],
                "morpheme": [
                    {"id": 1, "form": "줄다리기", "label": "NNG", "word_id": 1, "position": 1},
                    {"id": 2, "form": "는", "label": "JX", "word_id": 1, "position": 2},
                    {"id": 3, "form": "1900", "label": "SN", "word_id": 2, "position": 1},
                    {"id": 4, "form": "년", "label": "NNB", "word_id": 2, "position": 2},
                    {"id": 5, "form": "제", "label": "XPN", "word_id": 3, "position": 1},
                    {"id": 6, "form": "2", "label": "SN", "word_id": 3, "position": 2},
                    {"id": 7, "form": "회", "label": "NNB", "word_id": 3, "position": 3},
                    {"id": 8, "form": "파리", "label": "NNP", "word_id": 4, "position": 1},
                    {"id": 9, "form": "올림픽", "label": "NNG", "word_id": 4, "position": 2},
                    {"id": 10, "form": "부터", "label": "JX", "word_id": 4, "position": 3},
                    {"id": 11, "form": "1920", "label": "SN", "word_id": 5, "position": 1},
                    {"id": 12, "form": "년", "label": "NNB", "word_id": 5, "position": 2},
                    {"id": 13, "form": "제", "label": "XPN", "word_id": 6, "position": 1},
                    {"id": 14, "form": "7", "label": "SN", "word_id": 6, "position": 2},
                    {"id": 15, "form": "회", "label": "NNB", "word_id": 6, "position": 3},
                    {"id": 16, "form": "앤트워프", "label": "NNP", "word_id": 7, "position": 1},
                    {"id": 17, "form": "올림픽", "label": "NNG", "word_id": 7, "position": 2},
                    {"id": 18, "form": "까지", "label": "JX", "word_id": 7, "position": 3},
                    {"id": 19, "form": "정식", "label": "NNG", "word_id": 8, "position": 1},
                    {"id": 20, "form": "종목", "label": "NNG", "word_id": 9, "position": 1},
                    {"id": 21, "form": "이", "label": "VCP", "word_id": 9, "position": 2},
                    {"id": 22, "form": "었", "label": "EP", "word_id": 9, "position": 3},
                    {"id": 23, "form": "다", "label": "EF", "word_id": 9, "position": 4},
                    {"id": 24, "form": ".", "label": "SF", "word_id": 9, "position": 5}
                ],
            }
        ]
    }
  ]
}
```

부록 3: JSON 변환 시 구어 말뭉치의 마크업 기호 처리

1. 구어 말뭉치 마크업 기호 목록

<trunc> </trunc>: 끊어진 단어
<unclear> </unclear>: 잘 들리지 않아 추정하여 전사한 경우
<unclear/>: 잘 들리지 않아 전사를 못한 경우
<vocal desc="laughing"/>: 웃음
<vocal desc="목청가다듬는소리"/>: 목청
<vocal desc="applauding"/>: 박수
<vocal desc="singing"/>: 노래
<anon type="name"/>: 이름(비식별화)
<anon type="social-security-num"/>: 주민등록번호(비식별화)
<anon type="card-num"/>: 신용카드번호(비식별화)
<anon type="tel-num"/>: 전화번호(비식별화)
<note></note>

2. 기호 처리 방안

2.1. <trunc> </trunc>

- <trunc> </trunc> 사이의 요소는 분석 대상으로 삼는다.
- JSON 변환 시 <trunc>, </trunc>는 문자수에 포함하지 않고 begin, end 값을 잡는다.

원문: <u who="P2" n="49"><trunc>사</trunc> 싸우고 다투고 이러는데</u>

분석 대상: 사 싸우고 다투고 이러는데

JSON(word만 제시함.)


```
"sentence": [
    {
        "id": "SRDY00010.1.1.49", * id는 파일명.1.1.<u>의 n으로 부여
        "form": "사 싸우고 다투고 이러는데",
        "word": [
            {"id": 1, "form": "사", "begin": 0, "end": 1},
            {"id": 2, "form": "싸우고", "begin": 2, "end": 5},
            {"id": 3, "form": "다투고", "begin": 6, "end": 9},
            {"id": 4, "form": "이러는데", "begin": 10, "end": 14}
        ]
    }
]
```

2.2. <unclear> </unclear>

- <unclear> </unclear> 사이의 요소는 분석 대상으로 삼는다.
- JSON 변환 시 <unclear>, </unclear>는 문자수에 포함하지 않고 begin, end 값을 잡는다.

원문: <u who="P1" n="2617">인제 <unclear>오육십이면</unclear> 할 것 같애.</u>

분석 대상: 인제 오육십이면 할 것 같애.

JSON(word만 제시함.)


```
"sentence": [
    {
        "id": "SRDY00010.1.1.2617",
        "form": "인제 오육십이면 할 것 같애.",
        "word": [
            {"id": 1, "form": "인제", "begin": 0, "end": 2},
            {"id": 2, "form": "오육십이면", "begin": 3, "end": 8},
            {"id": 3, "form": "할", "begin": 9, "end": 10},
            {"id": 4, "form": "것", "begin": 11, "end": 12},
            {"id": 5, "form": "같애.", "begin": 13, "end": 16}
        ]
    }
]
```

2.3. <unclear/>

- JSON 변환 시 <unclear/>는 문자수에 포함하지 않고 begin, end 값을 잡는다.

원문: <u who="P2" n="2378">아 <unclear/> 있고</u>

분석 대상: 아 있고

JSON(word만 제시함.)


```
"sentence": [
    {
        "id": "SRDY00010.1.1.2378",
        "form": "아 있고",
        "word": [
            {"id": 1, "form": "아", "begin": 0, "end": 1},
            {"id": 2, "form": "있고", "begin": 2, "end": 4}
        ]
    }
]
```

2.4. <vocal>

- JSON 변환 시 <vocal>은 문자수에 포함하지 않고 begin, end 값을 잡는다.

원문: <u who="P2" n="263"><vocal desc="laughing"/> 살고 있어요.</u>

분석 대상: 살고 있어요. *** <vocal> 태그 뒤 공백 포함**

JSON(word만 제시함.)


```
"sentence": [
    {
        "id": "SRDY00010.1.1.263",
        "form": " 살고 있어요.",
        "word": [
            {"id": 1, "form": "살고", "begin": 1, "end": 3},
            {"id": 2, "form": "있어요.", "begin": 4, "end": 8}
        ]
    }
]
```

2.5. <anon/>

- <anon/>은 type과 n을 결합한 문자열을 문자수에 포함하여 begin, end 값을 잡는다.

원문: <u who="P2" n="2690">강남 <anon type="name" n="3"/> 안과요?</u>

분석 대상: 강남 name3 안과요?

JSON(word만 제시함.)


```
"sentence": [
    {
        "id": "SRDY00010.1.1.2690",
        "form": "강남 name3 안과요?",
        "word": [
            {"id": 1, "form": "강남", "begin": 0, "end": 2},
            {"id": 2, "form": "name3", "begin": 3, "end": 8},
            {"id": 3, "form": "안과요?", "begin": 9, "end": 13}
        ]
    }
]
```


2.6. <note>

- <u> 태그 이외의 요소이므로 분석 대상에 포함하지 않는다.

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2019년 12월 9일

발행일: 2019년 12월 9일
인 쇄: 성신POD

※ 이 책은 국립국어원의 용역비로 수행한 '형태 분석 말뭉치 구축' 사업의 결과물을 발간한 것입니다.

국립국어원
2019-01-23

형태 분석 말뭉치 구축

국립국어원


국립국어원